# From signal to metabolism: A journey through the regulatory layers of the cell

**D I S S E R T A T I O N**
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
**M.Sc. Timo Lubitz**

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin
Prof. Dr. Richard Lucius

Gutachter/innen

1. Prof. Dr. Dr. Edda Klipp
2. Prof. Dr. Hermann-Georg Holzhütter
3. Prof. Dr. Nils Blüthgen

Tag der mündlichen Prüfung: 3.5.2016

©2016 – Timo Lubitz
all rights reserved.

# List of publications and manuscripts included in this thesis

"Network reconstruction and validation of the Snf1/AMPK pathway in baker's yeast based on a comprehensive literature review" - **T Lubitz**, N Welkenhuysen, S Shashkova, L Bendrioua, S Hohmann, E Klipp, M Krantz, NPJ Systems Biology and Applications, volume 32 (pp. S238–S238), 2015

"Glucose derepression by yeast AMP-activated protein kinase SNF1 is controlled via at least two independent steps" - R García-Salcedo, **T Lubitz**, G Beltran, K Elbing, Y Tian, S Frey, O Wolkenhauer, M Krantz, E Klipp, S Hohmann, FEBS Journal 281 (7), 1901-1917, 2014

"New types of experimental data shape the use of enzyme kinetics for dynamic network modeling" - K Tummler*, **T Lubitz***, M Schelker, E Klipp, FEBS Journal 281 (2), 549-571, 2014

"Systematic construction of kinetic models from genome-scale metabolic networks" - NJ Stanford*, **T Lubitz***, K Smallbone, E Klipp, P Mendes, W Liebermeister, PLOS ONE 8(11): e79195, 2013

"Carbon tracking in colorectal cancer cells reveals novel influences of the KRAS and BRAF oncogenes on metabolic reprogramming" - R Fritsche-Guenther, S Kempa, E Klipp, **T Lubitz** (in alphabetical order), (*preliminary manuscript*)

"SBtab: An SBML-interconvertible table format for data exchange in Systems Biology" - **T Lubitz**, J Hahn, FT Bergmann, E Noor, E Klipp, W Liebermeister, Bioinformatics Journal, 2016 (*currently under revision*)

*shared first authorship

# List of publications and manuscripts not included in this thesis

”Annotation and merging of SBML models with semanticSBML” - F Krause, J Uhlendorf, **T Lubitz**, M Schulz, E Klipp, W Liebermeister, Bioinformatics 26 (3), 421-422, 2010

”Parameter Balancing in Kinetic Models of Cell Metabolism” - **T Lubitz**, M Schulz, E Klipp, W Liebermeister, The Journal of Physical Chemistry B 114 (49), 16298-16303, 2010

# Preface

This cumulative thesis describes how cell metabolism is influenced by the cells' reception of extracellular signals and how the diverse subsystems of this response can be modelled with mathematical approaches of Systems Biology. It is divided into four major parts, which describe the modelling of cell signalling pathways (Chapter 2.1), the modelling of cell metabolism (Chapter 2.2), a modelling approach for the analysis of metabolic reprogramming in a human cancer cell (Chapter 2.3), and the development of a standardised table format for Systems Biology and corresponding webservices (Chapter 2.4).

The first part, "Signalling pathways: From signal to transcription", describes approaches for modelling the transduction of extracellular signals in yeast cells and how these signals affect gene regulation. It comprises two publications: (i) "**Network reconstruction and validation of the Snf1/AMPK pathway in baker's yeast based on a comprehensive literature review**" was published in *NPJ Systems Biology and Applications* and my contribution was taking part in the extensive literature research, filling and maintaining the knowledge database, creating a Snf1 network reconstruction, generating a Boolean model of the network, and contributing to the writing of the manuscript. (ii) "**Glucose de-repression by yeast AMP-activated protein kinase SNF1 is controlled via at least two independent steps**" was published in *FEBS Journal* and my contribution was the generation of 24 different hypotheses models and a subsequent model discrimination process on the grounds of experimental data. I also worked on the manuscript.

The second part, "Enzymatic regulation of metabolic function", deals with the obstacles of and requirements for metabolic modelling. Furthermore, it proposes techniques for analysing the enzymatic regulation of metabolic models. This part comprises two publications: (i) "**New types of experimental data shape the use of enzyme kinetics for dynamic network modeling**" was published in *FEBS Journal*. For this review, my contribution was the conceptual design of presenting the development of enzyme kinetics and rate laws within the past 100 years. This description is followed by a state-of-the-art analysis of new experimental data and their influence on the usage of kinetic modelling today. Here, I wrote large parts of the manuscript. (ii) The second manuscript has been published in *PLOS ONE* with

the title "**Systematic construction of kinetic models from genome-scale metabolic networks**". My contribution to this work was taking part in the collaborative development of the workflow for large-scale metabolic modelling and providing the metabolic model with thermodynamically consistent parameter sets and convenience rate laws. I also contributed to the writing of the manuscript.

The third part, "Signals, genes, and metabolism: Towards a holistic view", introduces my work on cancer cell metabolism. It deals with metabolic reprogramming *via* oncogenes and is preliminarily titled "**Carbon tracking in colorectal cancer cells reveals novel influences of the KRAS and BRAF oncogenes on metabolic reprogramming**". My contribution was the creation of metabolic models and the execution of non-stationary metabolic flux analyses on them. I have also performed a goodness of fit approach to test the accuracy of the analyses and augmented the findings by consulting further experimental data. Finally, I wrote the preliminary manuscript on the project status.

The fourth and final part, "**SBtab: An SBML-interconvertible table format for data exchange in Systems Biology**", is a technical description of a standardised table format I helped to develop over the past years. This manuscript has been submitted to *Bioinformatics Journal* and is currently under revision. My contribution to this project was the joint development of the format, the implementation of software tools, the implementation of the corresponding web interface, and the writing of the manuscript.

The concluding words put the presented techniques and results into context of the state of the art and analyse their impact on our understanding of the biological systems.

Thesis advisor: Professor Edda Klipp Timo Lubitz

# From signal to metabolism: A journey through the regulatory layers of the cell


## Abstract

Cellular life is governed on different layers of regulation, which are tightly interconnected: (i) Signalling pathways transmit extracellular signals to the cells' nucleus, where (ii) gene regulation translates these signals into proteins, and (iii) proteins control metabolic functions, which convert nutrients to energy and cell building blocks. Due to the complexity of each of these systems, they are often analysed individually or only partially.

Systems Biology is an interdisciplinary field of research that offers techniques to harvest the information of todays high-throughput experiments. These techniques can be powerful approaches to investigate the aforementioned regulatory layers of a cell either individually or as a whole. In this thesis, I am employing means of Systems Biology to explore signalling pathways and metabolism, and I provide novel workflows for modelling and exploring these systems. Both workflows are focussed on accurate large-scale network reconstructions of the target system. Since one of the major problems in Systems Biology is the availability of experimental data, the workflows put emphasis on the handling of knowledge gaps. They are applied on the Snf1 pathway and metabolism in yeast and provide new findings about this model organism. Furthermore, this thesis presents an in-depth analysis of metabolic reprogramming in colorectal cancer cells, which yields previously unknown coherences of metabolic function and oncogenes. Finally, I am presenting a proposal for a standardised data format in Systems Biology, which is based on data tables.

In summary, this thesis comprises works on signalling pathways and cell metabolism, which includes novel modelling workflows and new biological findings, analyses their impact on the scientific state of the art, and proposes directions for new experimental targets.

## Zusammenfassung

Das Leben und Überleben einer Zelle wird auf verschiedenen Ebenen streng reguliert. Diese Ebenen sind eng miteinander verknüpft: (i) Signalwege leiten extrazelluläre Signale in den Zellkern, wo (ii) die Genregulation diese Signale zu Proteinen übersetzt, und (iii) Proteine kontrollieren metabolische Funktionen, die Nährstoffe zu Energie und zellulären Bausteinen konvertieren. Diese Systeme sind hochkomplex, so dass sie oft nur einzeln oder nur einzelne Teile davon betrachtet werden.

Systembiologie ist ein interdisziplinäres Forschungsgebiet, das Methoden anbietet, um die Informationen aus den heutigen Hochdurchsatz-Experimenttechnologien zu extrahieren. Diese Methoden können sehr effektiv sein, um die vorgenannten Systeme einzeln oder im Ganzen zu untersuchen. In dieser Doktorarbeit wende ich Methoden der Systembiologie an, um Signalwege und Zellmetabolismus zu erforschen, und ich präsentiere neue Arbeitsabläufe für das Modellieren und Analysieren dieser Systeme. Beide Methoden sind auf großskalige Netzwerkrekonstruktionen fokussiert. Da die Erhältlichkeit von experimentellen Daten eines der größten Probleme der Systembiologie darstellt, befassen sich die Methoden explizit mit dem Umgang mit Wissenslücken. Sie werden auf den Snf1 Signalweg und den Metabolismus von Hefezellen angewendet und vermitteln neue Erkenntnisse über diesen Modellorganismus. Des Weiteren präsentiert diese Arbeit eine eingehende Analyse vom metabolischen Reprogrammieren in Darmkrebszellen, welche bisher unbekannte Zusammenhänge von metabolischer Funktionalität und Onkogenen beinhaltet. Zum Abschluss stelle ich unseren Vorschlag für ein standardisiertes Datenaustauschformat vor, welches seinen Schwerpunkt auf Datentabellen der Systembiologie legt.

Zusammenfassend behandelt diese Doktorarbeit die Signalwege und den Metabolismus von Zellen, inklusive neuer Modellierabläufe und biologischer Erkenntnisse. Diese Erkenntnisse werden in den Kontext unseres aktuellen Wissensstandes gesetzt und darauf aufbauend werden neue potentielle Ansatzpunkte für Experimente vorgeschlagen.

# Contents

# Listing of figures

# Acknowledgements

I want to thank my supervisor **Edda Klipp** for her support and guidance. She did not only support me on a professional, but also on a personal level with her constant advise whenever it was required. She introduced me to researchers around the world and enabled me to develop independence in my research projects.

Next, I am very grateful to my inofficial mentors **Wolfram Liebermeister** and **Marcus Krantz**, who did not only improve my scientific advances significantly, but also taught me much about the secret politics of science and how to cope with repeated setbacks and obstacles.

As I am a person craving for harmony, I would not have been able to come this far without having my best friends also as colleagues. I consider myself a lucky person to have had the opportunity working besides **Katharina Albers**, **Marvin Schulz**, **Jannis Uhlendorf**, and **Max Flöttmann**. You guys are great people and I am grateful for knowing you. Not to mention that you also qualified as splendid proofreaders of this thesis. If I were a cell, you guys would be my favourite nutrition, glucose.

But of course, also the rest of the **group of Theoretical Biophysics** was a great support. I could work in an overwhelmingly friendly environment and was always able to find help in need. I want to highlight **Jens Hahn** for sitting down under the Christmas tree to proofread this thesis.

This list would not be complete without emphasising my family: My loving mother **Michaela**, my late father **Rainer**, my dear brother **George**, and my late sister **Karen** for making me who I am today. Last and with most passion I thank my loving wife **Anne** for her patience, her work, and her commitment. She deserves utmost gratitude for helping me with the figures of this thesis and the hurdles of this life. Work can be much easier with a person like her, always standing behind me, cuddled in a blanket, sipping a cup of tea, and patiently listening to my endless contemplations on the regulatory subsystems of the cell.

This thesis is dedicated to my little daughter Ella. She only knows about ten words so far, but often these are the exact words that I long to hear after a long day at work.

# 1 Introduction

## 1.1 The regulation of cellular life

Cells are the building blocks of all known life forms and understanding them is an important prerequisite to understand life itself. Some organisms only consist of a single cell, while others are vast conglomerates of different cell types forming various cell tissues; a human being, for instance, comprises approximately $10^{14}$ cells[1]. But although single cells are differing largely from each other, they are not as diverse as it might seem on first glance. Many cellular key features are highly conserved throughout different cell types and organisms. This allows us to conduct experiments on smaller cells - which are easier to culture and more simple in structure - and carefully transfer the results for a better understanding of complex human cellular systems[2]: Small organisms like the bacterium *Escherichia coli*, the fruit fly *Drosophila melanogaster*, the baker's yeast *Saccharomyces cerevisiae*, and many others have become model organisms due to their short generation times, the easy accessibility and manipulation of their genetic information, and other favourable cultivation features[3]. This makes them ideal living tools to explore cellular life[4].

A cell can be considered one large functional system that consists of many diverse regulatory subsystems. Three of the most important cellular subsystems are the following:

**Cell signalling** Cells need to be able to react to intracellular and extracellular signals in order to survive changing environmental conditions. Here, the term 'signals' is multifarious: It can describe extracellular salt concentrations, nutritient availability, ultraviolet radiation, pheromone concentrations, or pH value of the surrounding medium, but also intracellular signals like the output of DNA damage checkpoints during cell cycle. These signals are transmitted *via* intracellular signalling cascades that commonly target the cells' nucleus. Within the nucleus, the expression of target genes can be either inhibited or stimulated, which enables the cell to react to the source signal by adapting its protein composition[5,6].

**Metabolism** Cell metabolism transforms nutrition to energy and cellular building blocks. In a series of enzymatically catalysed reactions, carbon sources are broken into pieces (catabolism). On the one hand, these processes can be used to produce energy which is required for cellular activities. On the other hand, the broken down pieces of the nutritional source are used to build up new cellular components like lipids, nucleotides, or amino acids (anabolism). The provision of these components by metabolism drives the growth of the cell. It depends on the current status of the cell, whether catabolic or anabolic processes are preferred. This status is determined by intracellular and extracellular signals[7,8].

**Genetic regulation** The adaptation of a cell to environmental signals is mostly carried out by alterations in gene expression. An ample amount of transcription factors administers gene regulation, either inhibiting or stimulating gene expression in dependence of the cells' prevailing needs. The genetic information is transcribed to mRNA, which then is translated into proteins. Proteins operate all crucial cellular processes and thus execute the cellular response to signals[9,10].

These subsystems are only some of the many layers of cell regulation. They are complemented by complex mechanisms for cell cycle, transport mechanisms, volume regulation, cytoskeleton maintenance, and other pivotal processes. The analysis of these single modules in an either experimental or computational manner is a tedious task: Not only are they complex in themselves, they also are interconnected on many levels. The question arises what means are adequate to investigate the complex nature of these processes[11,12].

## 1.2 Systems Biology

The aforementioned biological systems and subsystems can be analysed with the approaches of a rather new field of research - Systems Biology. The birth of this field originates in the explosion of scientific knowledge acquisition and the development of new experimental techniques throughout the 20th century. This began with the first formulations of more sophisticated enzyme kinetics[13], peaked over milestones such as the construction of the Rutherford-Bohr atomic model[14], the discovery of the DNA double helix[15], the cloning of complex organisms like sheep[16], and ended on nothing less than the sequencing of the human genome[17]. The beginning of the 21st century then rather seemed to accelerate the speed of scientific evolution than to slow it down: With the dawning of the age of the internet, vast online databases[18,19,20] were filled with large amounts of experimental data processed by new high throughput experiments, and international collaborations among scientists were made possible by a few clicks of a mouse. But it was difficult to reasonably utilise these vast amounts of data and harvest the inherent information. These circumstances called for the rise of a new research field that could keep up with the dwindling speed of biological sciences. Its name is Systems Biology.

Systems Biology is an interdisciplinary field of research that combines methods and knowledge from computer science, mathematics, physics, chemistry, and biology. By shifting the focus from the previous reductionist view to a wider angle, Systems Biology is a holistic approach that tries to understand the 'big picture' of a system. A cell can be considered to be such a system and its overall behaviour is a property that emerges from the interplay of its subsystems: The system we want to explore is much larger than only the sum of its parts. Its subsystems are all interconnected, but not neccessarily only in an unidirectional way. For instance, on the one hand, genes are the origin of a complete organism[9], but, on the other hand, gene products (proteins) can in turn affect gene expression. Already in the 1960s, scientists practised this holistic approach[21], but it has become more and more popular and feasible with the increase of available techniques and data[22]. The ever-growing amounts of experimental data retrieved by microarrays, mass spectrometry, new generation sequencing, or similarly powerful approaches need to be incorporated into mathematical models to make us able to understand them. Systems Biology encompasses these advances, thus becoming the field of research that is required to tackle today's challenges in biological systems. In summary, Systems Biology offers new approaches to address the following problems:

- Terabytes of deep sequencing or "omics" data make the requirement for computational analyses apparent. Advances in data storage and data analysis are already obtained.
- The more complex contemporary biology becomes, the more it renders the intuitive understanding of biological systems infeasible.

- Bridges between different research disciplines need to be built to exploit their potentials wherever needed.
- Computational power is still increasing quickly and can be used for analyses tools, if these know how to exploit it reasonably.

Systems Biology has the theoretical potential to tackle these problems[22], which makes it reasonable to have a more detailled look at how it works in practise.

**The iterative cycle of Systems Biology** Ideally, Systems Biology underlies an iterative cycle of building/refining abstract models of biological target systems and iteratively conducting new experiments (see Figure 1.1). In the beginning, there is a hypothesis (or a question) about an existing biological system. To test this hypothesis, a model is built by employing the available data and knowledge about the system. The model construction is followed by computational simulations and analyses. The results of the latter can now initiate new iterative cycles for the successive improvement of the model by the proposal of new experiments and using the outcome as new model input. Thus, knowledge gaps can be identified and filled. Finally, the initial hypothesis should be verified or falsified. This proposed cycle is a generalisation of an idealised workflow in Systems Biology[23,24]. A closer look shows that, in practise, there are many more detailled and exhaustive workflows or pipelines for model creation, data incorporation, experimental or computational data generation, model analyses, and combinations thereof[25,26,27].

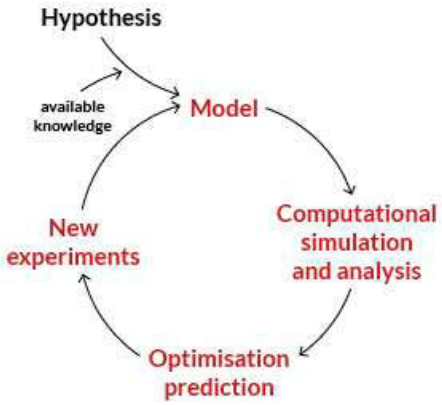



**Figure 1.1: The cycle of Systems Biology.** The cycle starts with a hypothesis about a biological system. By integrating the available knowledge and data, a mathematical model is created as an abstract representation of the system. It is used to perform computational simulations and analyses. Optimisations strive towards a more realistic system behaviour. These outputs are employed to conduct new experiments. With new data the model can be refined and another cycle starts. Ideally, the initial hypothesis is verified or falsified.

**Mathematical modelling approaches** A model of Systems Biology is supposed to describe the biological phenomena of a studied system by deploying mathematical formulae and techniques. In a best case scenario, a model reflects the behaviour of the corresponding system perfectly, but in practise this is highly unrealistic. Mathematical models are always an abstraction of reality. The role of mathematics in sciences has been elucidated by Box *et al.* in 1976: The employment of mathematics to subjects such as physics or statistics implies simplifying assumptions of which we know that they are wrong. The physicist knows that particles have mass, but still certain approximations rely on the assumption that they do not. Similarly, the statistician knows that normal distributions are extremely rare in nature, but still he often uses it for his analyses. Thus, we cannot expect a mathematical model that describes a biological system to be right, but we can say that it is fit to make proper predictions under certain conditions. Essentially, all models are wrong, but some are useful[28].

These observations make it even more important to thoroughly choose a modelling formalism that is suited for best describing the target system. Several criteria and questions come into play when a modeller chooses the formalism: (i) what kind of question is the model supposed to answer, (ii) how much experimental data is available for the description of its behaviour, (iii) should it be a small model with high detail or a coarse-grained model, and (iv) how good is the system explored so far (availability of knowledge)? Ideally, a model should represent the current knowledge about a system in an abstract and usable format[22]. The most common choices of model formalisms, which all have their advantages and disadvantages, are:

**Boolean models** Boolean models are a coarse-grained model formalism often used for large networks. The nodes of the network are switched either on or off, meaning activity or inactivity of the represented entities. They are connected by edges, which stand for the interaction behaviour of the connected nodes. Upon simulation, the node status are updated *via* the edges by Boolean update rules over discrete time steps. This formalism is conveniently used for gene regulatory networks or signalling networks (see Chapter 2.1) and does not need exhausting amounts of parameters. In general, Boolean models have a very crude time concept that is reduced to equitemporal simulation steps[29]. But they can be extended to include uncertainties about the node states, which makes them probabilistic Boolean models[30].

**Ordinary differential equations (ODEs)** ODE models represent each system observable with an ordinary differential equation. These allow a dynamic continuous simulation of a system, which results in the concentration changes of the observables over time. For this, they have to make several assumptions, like large and well-mixed molecule numbers and the choice of kinetic rate laws for the reactions. The required parameters can be estimated by the usage of experimental data and a plentitude of available software tools.

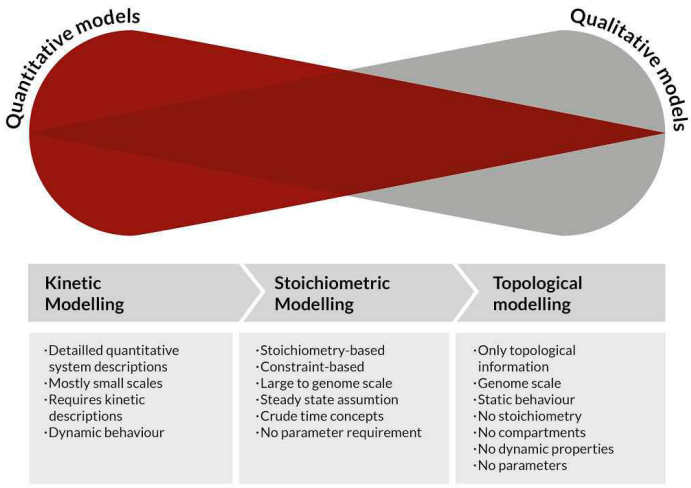



**Figure 1.2: Mathematical modelling: quality *versus* quantity.** Quantitative models show a high level of detail and allow for dynamic simulations and quantitative predictions. Their disadvantage is the high requirement of parameters that limits the possible model size. In comparison, qualitative models are large in size, but do not offer dynamic detail. In between are stoichiometric and constraint-based formalisms that aim a compromise between quality and quantity.

However, this prerequisite makes the usage of ODEs complicated for large systems. In regard of the systems' scope, ODEs can be applied to various kinds of biological target systems, e.g. signalling cascades (see Chapter 2.1.2), metabolic networks (see Chapter 2.2.2), or gene regulatory models[31,32].

**Stochastic models** Stochastic models are a counterpart to deterministic models like ODE models. The dynamics of the system are considered to be stochastic, which implies that the consideration of statistical physics needs to be employed for the analysis of such a model. This method can be applied e.g. for a system that involves few molecules of different substances and the individual molecules need to be traced. For further details about stochastic modelling the reader is referred to Wilkinson *et al.*[33].

**Stoichiometric models** For large metabolic models it can be useful to determine the flux distribution by analysis methods like flux balance analysis (FBA) or metabolic flux analysis (MFA). These approaches only require the stoichiometric information of a model and are not hampered by the need of kinetic parametrisation. Both applications can be improved by making them underlie individual constraints and both are available for steady state and dynamic (transient) analyses of the system (see Chapter 2.3)[34,35].

There are more modelling formalisms and also hybrids of multiple formalisms. An overview of the formalisms with respect to their detail and size is shown in

Figure 1.2. An extensive summary can be found in Machado *et al.*[36]. In this thesis, I have employed ODE models, Boolean models, and stoichiometric models.

**The exchange and reproducibility of data and models in Systems Biology** Mathematical models and experimental data should be formulated and stored in a reproducible manner. As aforementioned, Systems Biology aims at interconnecting the subsystems of a larger system to achieve a holistic view on it. But this aim requires the possibility to connect models with each other, incorporate experimental data into automated workflows and softwares, or visualise their content in an intuitive way. These are difficult tasks, since modelling formalisms and formats can be heterogenous; experimental data sheets are often layouted incidental or initially unintuitive; network visualisations tend to be unclear at first sight. All these problems can be tackled by the usage of standard data formats[37].

Standard formats are sets of guidelines and conventions for the representation of different kinds of information. Mathematical models can be stored in the widespread Systems Biology Markup Language (SBML ([38])) or MATLAB[39]. The Investigation/Study/Assay format (ISA-TAB) provides a spreadsheet format for the exchange of experimental data and experiment descriptions[40]. The Systems Biology Graphical Notation (SBGN[41]) offers three different languages for standardised graphical representations of biological networks. In interdisciplinary fields such as Systems Biology, there is a high demand for such exchange formats to facilitate the collaboration between experimentalists and modellers. They do not only enable an unproblematic data exchange, they also allow for the usage of diverse software tools (e.g. COPASI[42] or MATLAB for SBML models): If data adhere to standard formats, this facilitates automatic processing by software, which makes research results reproducible. Standards are also important in workflows: Standardised models and data can be easily routed through the various creation and analyses steps of exhaustive pipelines. Thus, standards ensure a high reusability of data, which prevents experiment repetitions or loss of information. This plethora of advantages advocates strongly for the usage of standard formats[43,37].

## 1.3 Structure and Outline

In my thesis, I am focussing on the adaptation of cell metabolism in response to extracellular signals. Firstly, these adaptations are crucial, since they ensure cell survival in changing environmental conditions. Secondly, understanding how cells respond to signals means a first step to finding drug targets for the treatment of diseases. But following the route of a signal from the extracellular space to actual changes in metabolic function requires the traversal of heterogenous regulation systems: A signal is received by a cellular receptor, transmitted *via* complex signalling cascades, and yields changes in gene expression. The resulting change in the protein composition of the cell is directly affecting enzyme levels and thus

the regulation of metabolic function. During this journey from signal to metabolism, several questions need to be addressed. What are appropriate modelling approaches for the very diverse systems? Are workflows for model generation at hand or do they have to be invented? What kind of experimental data is available and how can it be incorporated? How can we reasonably deal with knowledge gaps? And finally, - equipped with experimental data and Systems Biology modelling approaches - what can I contribute to our understanding of metabolic adaptations to diverse signals and how can this improve the treatment of diseases?

**Chapter 2.1** commences the initially announced route by focussing on the way from an extracellular signal over a signalling cascade to gene regulation. To traverse this first part of the overall route towards metabolism, modelling means for signalling pathways have to be employed. The question, how a signal can be transmitted to yield changes in gene regulation, needs to be addressed. For this purpose, I am introducing a novel workflow for the creation of signalling pathway reconstructions and their validation with Boolean modelling. This workflow is exemplified on the Snf1 pathway in yeast. Snf1 is the yeast homologue of mammalian AMPK and the pathway is responsible for glucose derepression of genes involved in the metabolism of alternative carbon sources. Next, I specifically address the main knowledge gap of this pathway, the activation mechanism of Snf1, with a model discrimination process based on experimental data. During these works on the Snf1 pathway, I built models of varying formalisms, identified knowledge gaps of the pathway, proposed solutions for the gaps, analysed the effect of the pathway on gene regulation, and introduced a general workflow for generating and validating signalling pathway reconstructions.

The subsequent **chapter 2.2** continues the journey from where the previous chapter left off: After an extracellular signal has induced changes in gene regulation, the protein composition of the cell is changed. This chapter deals with the effect that changes in enzyme concentrations have on the regulation of metabolic functionality. Firstly, and as a technical prerequisite, I present an extensive analysis of the evolution of enzyme kinetics and rate laws; these are mathematical formulations describing reaction dynamics. This preparatory excursion is followed by a proposed workflow for creating large-scale kinetic models of metabolism, which applies the previously reviewed aspects of enzyme regulation in practise. The workflow focusses on the dynamics and stability of kinetic metabolic models, which is strongly dependent on the parametrisation and enzymatic regulation. It is exemplified on a reconstruction of yeast metabolism and offers an intensive analysis on the regulatory principles of metabolic models.

After the previous chapters have completed the route from signal to metabolism piece by piece, **chapter 2.3** describes a more holistic view on this route. I am conducting a metabolic flux analysis of colorectal cancer cells, which is based on $^{13}C$ labelled isotope data and includes the experimental introduction of two mutated oncogenes. The analysis is extended by consulting proteomic and phosphoproteomic data acquired simultaneously from the same cell lines. This approach offers

a view on the regulatory subsystems of the cell in parallel and allows the drawing of conclusions about regulatory coherences.

**Chapter 2.4** is our proposal to a new standardised data exchange format - SBtab. It is based on common spreadsheet files and can thus be easily employed by scientists irregardless their technical prerequisites. Moreover, I have developed several tools for validating SBtab files and converting them to SBML. Besides some established standard formats of Systems Biology I have also used SBtab as exchange format in several workflows of my projects.

Finally, I am closing this thesis with concluding words about the possible impact of the proposed workflows and the introduced new biological findings.

# 2 Methods & Results

## 2.1 Signalling pathways: From signal to transcription

### 2.1.1 Reconstructing signalling pathways

A cell needs to be able to react to signals from its environment to ensure cell survival. Signals include nutrient availability, pheromone signals, salt concentrations, UV radiation, and many more. The transmission of such a signal is put into effect by a chain of intracellular reactions, which is called a signalling pathway. Different signals are often transmitted by different signalling pathways, but many of them are interconnected. The pathways yield changes in gene regulation, which represents how the cell reacts to the signal. This chapter focuses on the question, how signalling pathways can be modelled and what obstacles have to be faced in the process.

We propose a novel workflow for the generation of large-scale signalling network reconstructions. The reconstruction can be exported as a Boolean model, which enables a validation process for the network: A set of input nodes (representing stress signals) can be switched on or off, the model is then simulated, and the set of output nodes (gene regulation targets) is checked for conclusiveness. These input/output relationships are easily acquired literature data (e.g. "salt stress induces expression of the *ENA1* gene in yeast"[44]), and they can give a reasonable clue if the model transmits the signal realistically. If the input signals are not activating

the correct output nodes, the network needs to be scanned for loops and knowledge gaps.

I have exemplified the workflow on the Snf1 pathway in yeast, which is activated under glucose deprivation (and several other stresses) of the cell. By employing the workflow, I was able to identify several knowledge gaps, which were filled with hypotheticals to ensure signal transmission. After iterative rounds of model refinement, I could verify mechanistic connections between 4 input signals (glucose and nitrogen starvation, alkaline pH value, and salt stress) and 7 output nodes for transcriptional regulation. Several of these transcriptional outputs directly affect metabolism and can thus verify that cells under stress impose metabolic alterations via signalling cascades. This finding is by no means a revelation, but it is a proof of concept for the introduced workflow.

In summary, we could establish a connection from extracellular stress signals over the Snf1 pathway to genetic regulatory elements. The Boolean modelling approach overcomes the major obstacle of modelling signalling pathways, the lack of experimental data. Furthermore, we could point out, where gaps in the current knowledge about the Snf1 pathway lie and where we need to direct future experiments to close the gap between extracellular signals and gene regulation. Finally, we provide a novel workflow that enables researchers to apply the same analysis for other pathways of interest.

# Network reconstruction and validation of the Snf1/AMPK pathway in baker's yeast based on a comprehensive literature review

**Authors:** Timo Lubitz, Niek Welkenhuysen, Sviatlana Shashkova, Loubna Bendrioua, Stefan Hohmann, Edda Klipp, and Marcus Krantz

**Journal:** This article has been published in NPJ Systems Biology and Applications, 2015

**Accessiblity:** It can be accessed online via doi:10.1038/npjsba.2015.7

## Abstract

OBJECTIVES: The SNF1/AMPK protein kinase has a central role in energy homeostasis in eukaryotic cells. It is activated by energy depletion and stimulates processes leading to the production of ATP while it downregulates ATP-consuming processes. The yeast SNF1 complex is best known for its role in glucose derepression.

METHODS: We performed a network reconstruction of the Snf1 pathway based on a comprehensive literature review. The network was formalised in the rxncon language, and we used the rxncon toolbox for model validation and gap filling.

RESULTS: We present a machine-readable network definition that summarises the mechanistic knowledge of the Snf1 pathway. Furthermore, we used the known input/output relationships in the network to identify and fill gaps in the information transfer through the pathway, to produce a functional network model. Finally, we convert the functional network model into a rule-based model as a proof-of-principle.

CONCLUSIONS: The workflow presented here enables large scale reconstruction, validation and gap filling of signal transduction networks. It is analogous to but distinct from that established for metabolic networks. We demonstrate the workflow capabilities, and the direct link between the reconstruction and dynamic modelling, with the Snf1 network. This network is a distillation of the knowledge from all previous publications on the Snf1/AMPK pathway. The network is a knowledge resource for modellers and experimentalists alike, and a template for similar efforts in higher eukaryotes. Finally, we envisage the workflow as an instrumental tool for reconstruction of large signalling networks across Eukaryota.

### 2.1.2 Dealing with knowledge gaps of signalling pathways

Knowledge gaps in signalling pathways are large obstacles for every modelling effort. If a signal is not correctly transmitted throughout the model, the results will not be adequate. With the workflow introduced in the previous chapter I could identify new and verify old knowledge gaps in the yeast Snf1 pathway. One of the major issues is the mechanistic activation of Snf1 by glucose deprivation, a process that remains elusive. Step 5 of our workflow proposes means of dealing with such gaps: To ensure full connectivity of the reconstruction, hypotheticals can be added as place-holders to enable signal transmission. But this can only be a temporary solution until new experiments are directly targetted at this specific gap to offer new insight; an undertaking that is described in this chapter.

The activation of the Snf1 complex underlies a complex regulation. The involved components are protein kinases, phosphatases, and other regulatory elements. I have created 24 candidate ODE models, each representing a hypothesis for Snf1 complex activation. The models provide a time-resolved simulation of Snf1 activation and thus a more thorough insight into the system than a Boolean model. According to our experimental data and a discriminative parameter fitting, the most likely candidate model in accordance to the given data represents the following hypotheses:

The regulation of Snf1 phosphorylation is likely to be carried out by both kinase and phosphatase, not by one of them exclusively.

The phosphatase that regulates Snf1 dephosphorylation (protein phosphatase 1) is unlikely to be responsible for the localisation or dephosphorylation of the Snf1 main target suppressor, Mig1.

It seems highly likely that there is a second glucose-regulated step between the activation of Snf1 and the dephosphorylation (and thus inactivation) of repressor Mig1.

The conclusions drawn from the results mainly hint for the second glucose-regulated step that follows Snf1 activation. Until today, several hypotheses have been stated about this particular step, but they could not explain the mechanistic detail for Snf1 activation by glucose derepression. Nevertheless, with the presented results we have improved our knowledge about Snf1 activation, and this is a small but crucial step towards closing the gaps in the way from this extracellular signal to gene regulation changes.

# Glucose de-repression by yeast AMP-activated protein kinase SNF1 is controlled via at least two independent steps

**Authors:** Raul Garcıa-Salcedo, Timo Lubitz, Gemma Beltran, Karin Elbing, Ye Tian, Simone Frey, Olaf Wolkenhauer, Marcus Krantz, Edda Klipp, and Stefan Hohmann

**Journal:** This article has been published in FEBS Journal, 2014

**Accessiblity:** It can be accessed online via doi:10.1111/febs.12753

## Abstract

The AMP-activated protein kinase, AMPK, controls energy homeostasis in eukaryotic cells but little is known about the mechanisms governing the dynamics of its activation/deactivation. The yeast AMPK, SNF1, is activated in response to glucose depletion and mediates glucose de-repression by inactivating the transcriptional repressor Mig1. Here we show that overexpression of the Snf1-activating kinase Sak1 results, in the presence of glucose, in constitutive Snf1 activation without alleviating glucose repression. Co-overexpression of the regulatory subunit Reg1 of the Glc-Reg1 phosphatase complex partly restores glucose regulation of Snf1. We generated a set of 24 kinetic mathematical models based on dynamic data of Snf1 pathway activation and deactivation. The models that reproduced our experimental observations best featured (a) glucose regulation of both Snf1 phosphorylation and dephosphorylation, (b) determination of the Mig1 phosphorylation status in the absence of glucose by Snf1 activity only and (c) a regulatory step directing active Snf1 to Mig1 under glucose limitation. Hence it appears that glucose de-repression via Snf1-Mig1 is regulated by glucose via at least two independent steps: the control of activation of the Snf1 kinase and directing active Snf1 to inactivating its target Mig1.

## 2.2 Enzymatic regulation of metabolic function

### 2.2.1 The evolution of enzyme kinetics

In the previous chapter, we have established a modelling connection between extracellular signals and gene regulation. For this chapter, we are taking a resulting change in gene expression as prerequisite. Now that a signal has caused alterations in the gene expression of the cell, some protein levels are raised and others diminished. These changes have direct effects on cell metabolism: Some proteins impose allosteric regulation on cell compounds and thus cause changes in their reactivity. Others are enzymes with a direct effect on the rate of metabolic reactions. Now, the question arises how these regulatory effects can be modelled and analysed accurately.

The modelling of cell metabolism requires the consideration of numerous regulatory principles and implications. This includes different enzyme kinetics and the related difficult choice of kinetic rate laws for the modelled reactions; the incorporation of experimental data with emphasis on todays high throughput techniques; parameter acquisition and determination; and finally the means of validating the model with respect to parameter sensitivity, identifiability, and their confidence intervals. These technical prerequisites are reviewed in this chapter, and can be considered the foundation of the exhaustive metabolic modelling workflow of the next chapter.

# New types of experimental data shape the use of enzyme kinetics for dynamic network modeling

**Authors:** Katja Tummler*, Timo Lubitz*, Max Schelker, and Edda Klipp

**Journal:** This article has been published in FEBS Journal

**Accessiblity:** It can be accessed online via doi:10.1111/febs.12525

***** These authors contributed equally to this work

## Abstract

Since the publication of Leonor Michaelis and Maude Menten's paper on the reaction kinetics of the enzyme invertase in 1913, molecular biology has evolved tremendously. New measurement techniques allow *in vivo* characterization of the whole genome, proteome or transcriptome of cells, whereas the classical enzyme essay only allows determination of the two Michaelis–Menten parameters V and $K_m$ . Nevertheless, Michaelis–Menten kinetics are still commonly used, not only in the *in vitro* context of enzyme characterization but also as a rate law for enzymatic reactions in larger biochemical reaction networks. In this review, we give an overview of the historical development of kinetic rate laws originating from Michaelis–Menten kinetics over the past 100 years. Furthermore, we briefly summarize the experimental techniques used for the characterization of enzymes, and discuss web resources that systematically store kinetic parameters and related information. Finally, describe the novel opportunities that arise from using these data in dynamic mathematical modeling. In this framework, traditional *in vitro* approaches may be combined with modern genome-scale measurements to foster thorough understanding of the underlying complex mechanisms.

### 2.2.2 Creating large-scale kinetic models of metabolism

Properly calibrated kinetic network models are a practical tool to identify knowledge gaps *in silico* and direct subsequent experiments, which is a comparable approach to what was introduced in the previous chapter about signalling pathways. Although such a model is just an abstract representation of *in vivo* metabolism, it can well identify poorly characterised parts of the pathway. Besides the consideration of technical prerequisites in the previous chapter, also the construction of a kinetic metabolic model in practise is not trivial. With growing model complexity, kinetic parametrisation renders more and more infeasible: Too many parameters have not been determined by experiments, cannot be determined, or have been determined but are not available to the modeller. Thus, Systems Biology requires new workflows for the construction of large-scale kinetic models, which can on the one hand grow in size, but on the other hand do not lose in terms of sensitivity and confidence. This raises the importance of sophisticated approaches for data augmentation to deal with fragmentary kinetic parameter collections. Furthermore, adding enzymatic regulation to the model dynamics requires experimental data as well as a sound choice of rate laws to not corrupt the model's stability. This chapter is focused on how these diverse but interconnected problems can be addressed in order to construct a large-scale dynamic model of metabolism.

We have established a workflow for large-scale kinetic modelling and introduced it on the example of yeast metabolism. The model structure and stoichiometry is extracted from a genome-scale model of yeast metabolism. Referring to the questions raised in the previous chapter, our workflow comprises: (i) the choice of a common modular rate law for the reactions rate laws (a generalised form of the reversible Michaelis-Menten kinetics applicable to any reaction stoichiometry), (ii) a parametrisation via collected literature data and data augmentation with parameter balancing (ensuring thermodynamic consistency), and finally (iii) a sensitivity and metabolic control analysis of the model to ensure a realistic behaviour under perturbations of extracellular nutrients and enzyme concentrations. The resulting model shows realistic metabolic fluxes and can be directed to a desired steady state, which is also stable in response to system perturbations. So far, there was no Systems Biology workflow ensuring these accomplishments for large-scale metabolic models. Finally, we figure that the step-by-step description of the workflow for yeast metabolism will enable its unproblematic application to other organisms.

# Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks

**Authors:** Natalie J. Stanford*, Timo Lubitz*, Kieran Smallbone, Edda Klipp, Pedro Mendes, and Wolfram Liebermeister

**Journal:** This article has been published in PLOS One, 2014

**Accessiblity:** It can be accessed online via doi:10.1371/journal.pone.0079195

* These authors contributed equally to this work

## Abstract

The quantitative effects of environmental and genetic perturbations on metabolism can be studied *in silico* using kinetic models. We present a strategy for large-scale model construction based on a logical layering of data such as reaction fluxes, metabolite concentrations, and kinetic constants. The resulting models contain realistic standard rate laws and plausible parameters, adhere to the laws of thermodynamics, and reproduce a predefined steady state. These features have not been simultaneously achieved by previous workflows. We demonstrate the advantages and limitations of the workflow by translating the yeast consensus metabolic network into a kinetic model. Despite crudely selected data, the model shows realistic control behaviour, a stable dynamic, and realistic response to perturbations in extracellular glucose concentrations. The paper concludes by outlining how new data can continuously be fed into the workflow and how iterative model building can assist in directing experiments.

## 2.3 Signals, genes, and metabolism: Towards a holistic view

The two routes from a signal to gene regulation and from gene expression to metabolism were the focuses of the previous chapters. But although I have introduced large-scale modelling approaches for both signalling pathways and metabolism, Systems Biology demands a more holistic view on the system. This can be achieved by taking into account different layers of regulation not subsequently, but simultaneously. First of all, this requires experimental approaches sophisticated enough to provide us with simultaneous data from different cellular subsystems. And even if this can be achieved, we need ideas how to incorporate these data into one conclusive mathematical model. This chapter deals with this exact scenario.

Within this project, we are elucidating metabolic reprogramming in colorectal cancer cells. It has been shown that metabolic features of cancer cells differ strongly to these of normal cells. To investigate these features, I am provided with a set of heterogenous experimental data types, i.e. $^{13}C$ label incorporation data of metabolic observables for different time points, label-free measurements of 4000 genes using shot-gun proteomics, and phosphorylation patterns of key players in signalling cascades *via* a BioPlex system. These experimental approaches, all performed on the same cancer cell lines in wild type and with mutations of two common oncogenes, are the basis for a complex analysis of the regulatory patterns of cancer cells. As a starting investigation, I have employed the labelled isotope data for nonstationary metabolic flux analyses (MFA) of glycolysis, TCA-cycle, and glutaminolysis for all used cell lines. This approach allows the determination of metabolic fluxes from transient isotope labelling experiments. The accuracy of these analyses was tested by assessing the goodness of fit of the models. The MFA results confirm known behaviours of cancer cells, as well as identify novel findings.

The most interesting findings show the dependence of certain metabolic behaviours on the KRAS protein. The MFA suggests that cells with a mutated KRAS oncogene fail to accumulate lactic acid, which is one of the key features of the Warburg effect in cancer cells. Furthermore, cells bearing this mutation route carbon from $^{13}C$ glutamine into the reductive part of the TCA cycle. These results are augmented by analysing the proteomic and phosphoproteomic data of the cell lines and by inferring connections between the subsystems. Understanding these underlying principles is a crucial step in the identification of potential drug targets. To augment our analysis beyond the capabilities of a metabolic flux analysis, we propose means of improving the experimental as well as the theoretical side of this project by increasing measurement observables and explicitly suggesting modelling techniques, which incorporate more of the provided data.

# Carbon tracking in colorectal cancer cells reveals novel influences of the KRAS and BRAF oncogenes on metabolic reprogramming

Raphaela Fritsche-Guenther[1], Stefan Kempa[1], Edda Klipp[2], Timo Lubitz[2]
(in alphabetical order)

January 3, 2016

1. Integrative Proteomics and Metabolomics Platform, Berlin Institute of Medical Systems Biology at the Max-Delbrück Center for Molecular, 13125 Berlin, Germany.
2. Humboldt-Universität zu Berlin, Institut für Biologie, Theoretische Biophysik, Invalidenstraße 42, D-10115 Berlin.

**Abstract**

The regulatory principles of cancer cells are differing in many aspects from those of healthy cells: Metabolic activity is reprogrammed by oncogenes and their direct and indirect influences on metabolic enzymes and metabolites. Cancer cells favour aerobic and anabolic processes, mostly independent of growth factor availability and cellular oxygen levels. Understanding these changed features is a key aspect to cancer treatment. We are approaching this aim by investigating the behaviour of colorectal cancer cells with cutting edge experimental techniques ($^{13}$C carbon tracing with the pSIRM workflow) and mathematical modelling approaches (metabolic flux analysis), which reveals novel causes of metabolic reprogramming. We suggest the KRAS protein to be responsible for lactic acid accumulation in cancer cells and we narrow down the search for responsible participants of this effect with proteomic and phosphoproteomic data. Furthermore, proteins BRAF and KRAS both appear to be responsible for the employment of the reductive TCA cycle in glutaminolysis. Finally, we are proposing targets for future experiments and further mathematical modelling approaches for cancer cell metabolism.

## Introduction

Colorectal cancer has the third highest frequency among tumors in developing countries with approximately one million new cases per year [1]. The disease arises from an imbalance in differentitation, proliferation, and apoptosis of the epithelium triggered by specific mutations. Up to 50% of patients harbor a mutation in the KRAS oncogene. Beside RAS activation of downstream signalling cascades, which are important for several cell processes, studies have shown the importance of metabolic changes induced by oncogenic KRAS [2, 3]: Metabolic deregulation is one of the new hallmarks in cancer shown by Hanahan and Weinberg [4, 5]. Several reports show that nutrient uptake and metabolic alterations are under control of the RAS protein, leading, for instance, to changes in glucose and glutamine consumption [6], increased glycolysis [7], and induction of *de novo* lipid synthesis [8, 9] (see Figure 1 for an overview of metabolic repro-

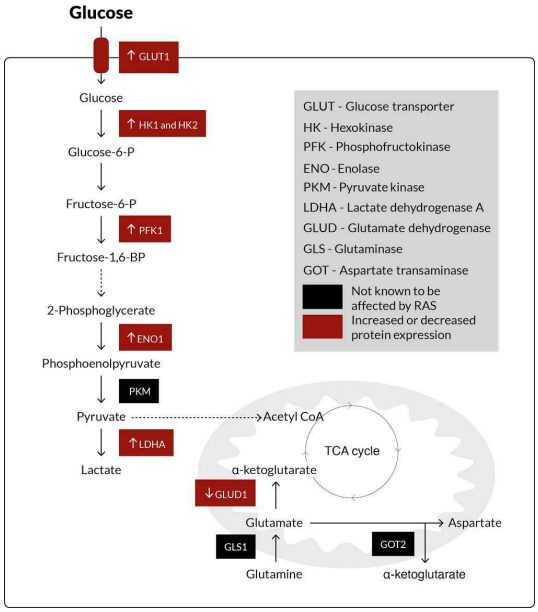



Figure 1: **Influences of the RAS oncogene on metabolism.** The RAS protein regulates many enzymatic targets in central carbon metabolism and is responsible for metabolic reprogramming in cancer cells. The depicted reactions are only an extract of RAS targets in central carbon metabolism. Dashed lines represent clustered reactions.

gramming induced by RAS in cancer cells). In addition to *KRAS* mutations, *BRAF* mutations occur in approximately 10% of colorectal cancer cells and are witnessed in almost all colorectal tumors with wildtype *KRAS*. Data implicates that around 50% of patients harboring no mutation in *KRAS* could benefit from an anti-EGFR therapy. Nevertheless, 40-60% of patients with wildtype *KRAS* do not respond to the treatment [1]. Advances in drug development over the last decades have expanded the number of potential therapies, but these urgently require optimisation. Therefore, understanding the underlying biology of how colorectal tumors form and progress is important for developing effective personalised therapies for patients with this disease.

We aimed at investigating cell morphology alterations upon stable expression of both oncogenes over time. Since oncogenes can induce a metabolic switch, we further analysed the effect of nutrient changes in cells expressing mutant BRAF and KRAS proteins. We used Caco-2 cells, which were stably transfected with inducible *BRAF* or *KRAS*. The Caco-2 cell line is a commonly used model for analysing human intestinal epithelium. The cells are known to be capable of spontaneous differentiation under standard conditions *in vitro* and harbor enterocyte-like structural and functional characteristics [10].

We are employing the recently introduced pSIRM workflow (pulsed stable isotope-resolved metabolomics [11]), which allows for the direct measurement of dynamic metabolic activity by tracing the flux of $^{13}$C glucose, $^{13}$C glutamine, or other labelled isotopes throughout the central carbon metabolism. The resulting mass isotopomer distribution vectors can be employed for an instationary metabolic flux analysis (MFA [12, 13]), which quantitatively determines the metabolic fluxes of the system (for an overview of the technique see Figure 2). The pSIRM approach is extended for shotgun proteomic sequencing and BioPlex determination of phosphoproteomics to investigate

the causes of metabolic deregulation. Employing these diverse data for a systematic analysis is a prerequisite for the identification of potential genetic targets for the optimisation of a specific phenotype [14] or for metabolic alterations upon the change of environmental conditions and genetic modifications [15, 16].

We were able to verify known and discover novel features about the connection of BRAF and KRAS proteins to lactic acid production, as well as to the occurrence of reductive TCA cycle in glutaminolysis. Still, we are far from understanding how the microenvironment modulates tumor heterogeneity and drives the phenotypic behaviour of a tumor cell population, but high throughput technologies, such as proteomics and metabolomics, aim at a global molecular description of complex cellular behaviour. Metabolic profiling in combination with sophisticated mathematical modelling approaches enables the comparison of normal and mutated cells, which can be used to identify new biomarkers and support for cancer diagnosis and treatment.

# Methods

**Cell culture** The Caco-2 cell lines were kindly provided by Dr. Tilmann Brummer (Institute of Molecular Medicine and Cell Research Freiburg). Caco-2tet cells and their derivates Caco-2tet/empty vector, Caco-2tet/BRAF$^{V600E}$, and Caco-2tet/KRAS$^{G12V}$ have been described previously [17, 18]. Also, the doxycycline inducible expression system is described in detail elsewhere [19]. The cells were incubated in a humified atmosphere of 5% CO in air at 37° Celsius and cultivated in glucose-free Dulbecco's modified Eagle's medium (DMEM, Life Technologies, #A14430-01), supplemented with 10% fetal bovine serum (GIBCO), and 1% penicillin/streptomycin (GIBCO), 4 mM glutamine (GIBCO), 1 g/l glucose, 2 $\mu$g/ml doxycycline, 5 $\mu$g/ml puromycin, and 5 $\mu$g/ml blasticidin. After seeding, the cells were cultured with a stable addition of doxycycline over time to induce KRAS/BRAF expression. Cells were preincubated for 3 days in 1g/l glucose, plated and lysed after 48 h.

**pSIRM** We employed the pSIRM workflow (pulsed stable isotope-resolved metabolomics [11]), to measure the dynamic metabolic activity through central carbon metabolism. The approach allows the measurement of time-resolved isotopic enrichment and a quantification of metabolites within a single measurement. Thus, the metabolites closer to the originating substrate will have a higher isotope label incorporation than those further downstream. The measured observables are glucose 6-phosphate, glyceraldehyde 3-phosphate, serine, pyruvic acid, lactic acid, alanine, citric acid, fumaric acid, and malic acid for [u$^{13}$C] glucose incorporation (at time points 2, 5, and 8 minutes). For measurements with [u$^{13}$C] glutamine, the observables are glutamic acid, $\alpha$-ketoglutaric acid, succinic acid, fumaric acid, malic acid, and the 276 and 278 isoforms of citric acid (at time points 5, 15, and 45 minutes). For a thorough method description the reader is referred to the original publication introducing the pSIRM workflow [11].

The approach has been extended for the simultaneous determination of label-free proteomic quantities by shotgun sequencing [20] and of phosphoproteomic data acquired by the BioPlex technology (data not shown).

**Instationary MFA** Metabolic flux analysis (MFA) is a fluxomics technique for estimating the metabolic fluxes of a system on the grounds of experimental data. Based upon stable isotope $^{13}$C data, MFA aims at quantifying the integrated responses of metabolic networks to system changes [12]. The approach can be ex-

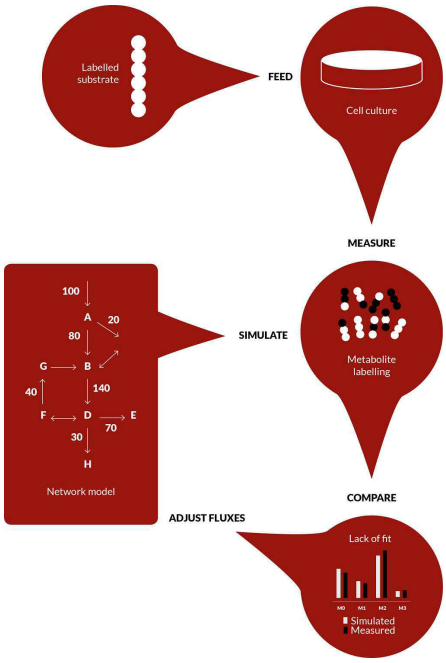



Figure 2: **Workflow for metabolic flux analyses** MFA studies are carried out by feeding cells an isotopically labelled substrate (here, $^{13}$C-labeled glucose and glutamine) and measuring the patterns of isotope incorporation that emerge in downstream metabolites using mass spectrometry. A computational model of the intracellular metabolic network is used to determine pathway fluxes by integrating these isotope labelling data. The results can be used for iterative rounds of experiments and simulations. Finally, MFA reconstructs comprehensive flux maps depicting cell metabolism.

tended by employing transient isotope labelling experiments, which yields an isotopically non-stationary MFA [13].

Mathematically, MFA is an optimisation search for the identification of flux parameters and pool sizes, which minimise the sum-of-squared residuals (SSR) between the experimentally derived data and the computationally simulated data [21]. At each iteration of the optimisation, the objective function is solved to simulate the measured isotopomer distributions on the basis of the metabolic network and a set of parameter estimates. The inverse problem can be expressed as

$$\max_{u,c,h} \Phi = r^{\mathsf{T}} \cdot \sum^{-1} \cdot r$$
$$\text{subject to } K \cdot u \geq 0, c \geq 0, h \geq 0.$$

To determine the minimal distance between computational simulation and experimental data, the objective function $\Phi$ needs to be maximised. First, the determination of the measurement vector m is required, which is a general function of the fluxes $u$, the pool sizes $c$, and time $t$. Thus, $m$ comprises all simulations of fluxes, pool sizes, and mass isotopomer distributions, for which experimental measurements are available. In a second step, by incorporating an MS scaling factor $h$ for the renormalisation of mass distribution vectors [22], the residual vector $r$ can be calculated from $m$ as

$$r = m(u, c, h, t) - \hat{m}(t), \tag{1}$$

where $\hat{m}(t)$ is the experimental measurement vector. Now, $\Phi$ can be maximised over $r$ and the diagonal weighing matrix $\sum^{-1} = diag(\sigma)$, yielding the optimal flux distribution for the metabolic network in terms of the experimental data. For detailled reviews on MFA the reader is referred to [21] and [23].

All models are available as MATLAB [24] model structure and flux map files.

**Goodness of Fit** After performing a metabolic flux analysis, it is of utmost importance to assess the goodness of the fit to evaluate the accuracy of the analysis. This test is based on the assumption that the SSR achieved by the objective function follows a $\chi^2$ distribution [25] with $n - p$ degrees of freedom (where $n$ is the number of independent experimental measurements and $p$ the number of parameters that are fitted). If the fit is rejected, either the model or data are likely to comprise inconsistencies that need to be reevaluated before proceeding. For a proper description of the method the reader is refered to Riedwyll et al. [26]. For the MFA and the goodness-of-fit analysis, we were using the MFA-INCA toolbox [27].

# Results

**KRAS protein induces lactic acid accumulation** The paths of the labelled carbon atoms after $^{13}$C glucose incorporation can be followed throughout glycolysis and the TCA cycle. The corresponding label incorporations are depicted in Figure 3 (A). Furthermore, Figure 3 (B) shows the metabolic models designed for the metabolic flux analysis: All measured metabolites are incorporated and the fluxes are enabled to branch into sinks. Figure 3 (C) displays the goodness of fit for the metabolic flux analyses.

The carbon routing pattern is similar for all three cell lines. As expected, the largest amounts of isotope labelled carbon is routed through glycolysis, without branching significantly into the serine synthesis branch. Over the branching point pyruvic acid, the carbon is distributed in different branches of metabolism: In both the control and Caco-2 BRAF$^{V600E}$ cell lines, about 50% of the introduced carbon branches into the production of lactic acid. Furthermore, large quantities are employed for the synthesis of alanine (18.99% in control and 31.88% in BRAF$^{V600E}$ cells).

Regarding the pyruvate branching point, Caco-2 KRAS$^{G12V}$ cells show significant differences to the other two cell lines. The production of lactic acid is decreased significantly, while the production of alanine precursors has increased to a similar degree at which lactic acid production is decreased. Finally, the rest of the labelled carbon is routed through the TCA cycle, where no measurable branches could be identified by the analysis. The assessment of the goodness of the fits (see Figure 3 (C)) verifies that the metabolic flux analysis results were accepted, although some outliers are situated around the simulation line of the weighing results.

**The reductive TCA cycle in glutaminolysis is employed by mutations in BRAF and KRAS** Cells fed with $^{13}$C glutamine substrates are routing the labelled carbon directly into glutaminolysis. Glutamine is converted to glutamic acid, which then enters the mitochondrion; the carbon can be tracked throughout the TCA cycle. We are able to determine whether the introduced carbon is routed into the reductive or the oxidative TCA cycle by examining the mass isotopomer distribution vector of citric acid: Depending on the positions of the labelled carbon atoms, it can stem (i) from glycolysis, (ii) the oxidative TCA cycle, or (iii) from the reductive TCA cycle. Since the introduced $^{13}$C glutamine excludes a significant amount of citric acid synthesised from glycolysis (i), we distinguished between the two citric acid isoforms stemming from oxidative (ii) and reductive TCA cycle (iii). Figure 4 depicts the experimentally derived label incorporations (A), the corresponding metabolic flux analyses (B), and the goodness of fit for the analyses (C) of the three different cell lines.

Similar to the glucose experiments, the label

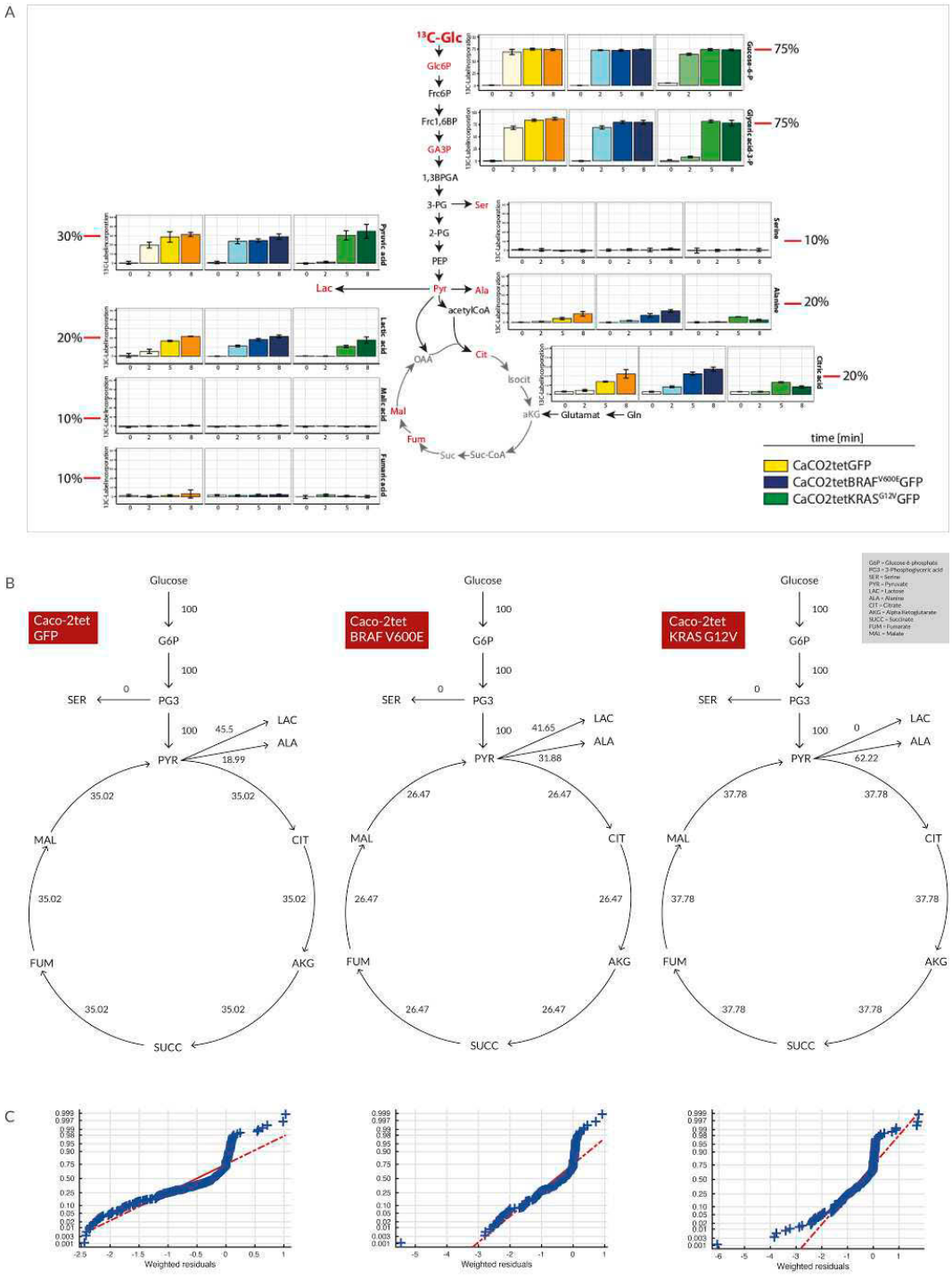



Figure 3: **Carbon tracking with $^{13}$C glucose.** (A) Experimentally derived label incorporation in central carbon metabolism after addition of $^{13}$C glucose. The different colours correspond to the used cell lines, the histograms depict the incorporation for the measured time points. (B) The metabolic fluxes as simulated by MFA. The fluxes are estimated on the grounds of experimental data, in this case $^{13}$C glucose isotope-resolved metabolomics. (C) Goodness of fit for the MFA. This technique gives an estimate about the accuracy of the corresponding metabolic flux analysis.

incorporation of the three cell lines is similar for most reaction routes. Glutamine is converted to glutamic acid and enters the TCA cycle *via* $\alpha$-ketoglutaric acid. The control cell line routes all carbon in the oxidative TCA cycle, but branches 50% of it out *via* fumaric acid. While this branch remains constant among all used cell lines, the Caco-2 BRAF$^{V600E}$ cells direct a small amount of carbon to the reductive TCA cycle (3.61%), and the Caco-2 KRAS$^{G12V}$ cells even 14.4%. The assessment of the goodness of fits, Figure 4 (C), allows the inference of conclusions about the accuracy of the analyses. While the quality of the fits as a whole is located within the acceptance boundaries, all three cell lines show some outliers in the bottom left of the fit visualisations.

## Discussion

**KRAS and BRAF are responsible for metabolic reprogramming** The examined cancer cell lines show characteristics of metabolic reprogramming dependent on the oncogenes KRAS and BRAF. Cancer cells are known to accumulate lactic acid as a compensation for increased glucose uptake [28]. Our experiments show the direct connection to an important player of growth factor signalling, the GTPase KRAS: In Caco-2 cells with a G12V mutation of the *KRAS* gene, lactic acid is not accumulating as in Caco-2 wild type or BRAF$^{V600E}$ cells. It can be concluded that a mutation of the *KRAS* gene either leads to severe effects on downstream signalling targets of KRAS protein or the mutated protein has a direct influence on metabolic function. Recent studies support the theory that activated oncogenes often reprogram cellular metabolism in a direct manner, instead of interfering with signalling cascades [4]. Moreover, our flux analyses show a high amount of carbon routed to the synthesis of alanine in all three cell lines, which has been shown in previous studies [29, 30, 31].

We were also able to analyse features of cellular glutaminolysis. When fed with $^{13}$C glutamine, all cell lines showed a high leakage of labelled carbon from oxidative TCA cycle *via* the branching point fumaric acid. This occurrence has been shown previously in cancer cells and is likely caused by mutations in the fumaric acid hydratase (FH) and the succinate dehydrogenase (SDH). This causes an accumulation of fumaric acid and succinic acid, which is compensated by the production of succinic acid glutathione [32, 33].

Finally, while the Caco-2 control cell line, provided with $^{13}$C glutamine, routes the labelled carbon entirely through the oxidative TCA cycle, Caco-2 BRAF$^{V600E}$ cells direct a small amount of it in the reductive cycle. This amount is even increased in Caco-2 KRAS$^{G12V}$ cells, which has been shown before for cancer cells [34]. The stepwise increased carbon amount in the two cell lines proposes a multicausal effect. While the BRAF$^{V600E}$ mutation already employs the reductive TCA cycle, this is amplified by a KRAS mutation. Similar to the lactic acid production in the other experiments, the various downstream targets of KRAS protein are potential contributors to this amplification from BRAF to KRAS mutant.

**Augmenting the results by proteomic and phosphoproteomic data** The metabolic fluxes give an accurate description of how cancer cell metabolism is reprogrammed in contrast to normal cells. However, while the reprogramming effects in cancer cells become more and more clear (see Figure 1), the dependency of these effects on oncogenes often remains elusive. These dependencies can only be studied accurately by simultaneous measurements of other regulatory cell structures. Thus, we are augmenting our findings so far towards a

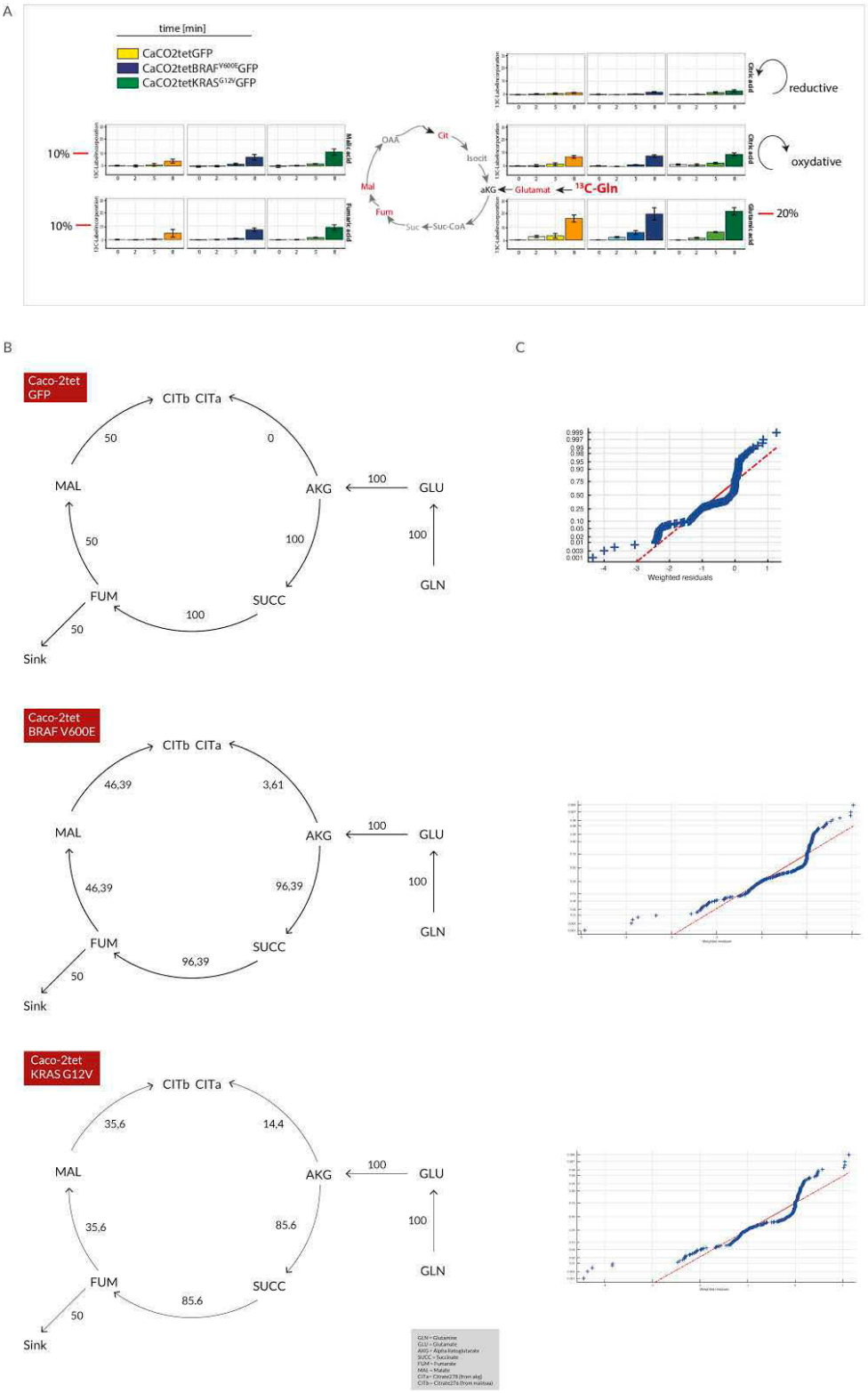



Figure 4: **Carbon tracking with $^{13}$C glutamine** (A) Experimentally derived label incorporation in the TCA cycle after addition of $^{13}$C glutamine. The different colours correspond to the used cell lines, the histograms are depict the incorporation for the measured time points. (B) The metabolic fluxes as simulated by MFA. The fluxes are estimated on the grounds of experimental data, in this case $^{13}$C glutamine isotope-resolved metabolomics. (C) Goodness of fit for the MFA. This technique gives an estimate about the accuracy of the corresponding metabolic flux analysis.

more holistic view on the regulatory principles of the system. The proteomics analysis (unpublished data not shown here) could identify genes *SLC2A3* and *SLC2A14*, members of the glucose transporter family GLUT, to be severely underexpressed in KRAS$^{G12V}$ cells in comparison to the other used cell lines, which coincides with earlier studies [35]. The enzyme lactate dehydrogenase A (LDHA), which is responsible for the conversion of pyruvate to lactic acid, showed decreased concentration levels in the KRAS$^{G12V}$ cells. This is in agreement with the decreased lactic acid concentrations in this cell line. There were no significant changes in the concentration levels of lactate dehydrogenase B (LDHB), which also underlines the hypothesis that certain cancer forms prefer different enzyme isoforms over others [36, 37]. Furthermore, this is contradictory to an effect witnessed in lung cancer cells, which show the linkage of LDHB to the KRAS oncogene [38].

The phosphoproteomic data (unpublished data not shown here) provides a view on the phosphorylation patterns of the downstream targets of KRAS protein. KRAS triggers a variety of signalling pathways, which are mostly transmitting growth factor signals. Our data exclude the RAS/MAPK pathway as potential cause of the lactic acid accumulation: Since the increased lactic acid production is not tempered in BRAF$^{V600E}$ cell lines, it is unlikely that the RAS/MAPK pathway, which employs BRAF protein, is responsible for the witnessed effect [39]. Other KRAS downstream targets include PI3K (cell survival, growth, diverse transcription factors), RALGDS (endocytosis), mTOR, Protein kinase B, and several other frequented compounds. The phosphorylation patterns revealed changes for Protein kinase B (and its downstream targets NF-$\kappa$B, mTOR, and p70S6). These alterations suggest that this signalling pathway is frequented differently than in the other cell lines, thus giving rise to the assumption that it is involved in the changed behaviours of the cell, i.e. decreased lactic acid production.

In the analysis of the reductive TCA cycle effects in glutaminolysis, we are again led to either direct or downstream targets of KRAS protein. But here, we cannot exclude the RAS/MAPK pathway, since the BRAF$^{V600E}$ cell lines already show carbon routing through the reductive TCA cycle. On the contrary, we must explicitly include downstream targets of the RAS/MAPK pathway, and extend this list by targets of the other KRAS-induced pathways. Intuitive targets are the enzymes solely employed for reductive TCA (ATP citric acid lyase, 2-oxoglutarate:ferredoxin oxidoreductase, and pyruvic acid:ferredoxin oxidoreductase). Here, we can exclude the ATP citric acid lyase, which is responsible for the synthesis of acetyl-CoA and oxaloacetate from citric acid. Our proteomics data suggest that there are no shifts in gene expression for this enzyme in our different cell lines.

**Directing future research** Our data highlight the complexity of the metabolic reprogramming of cancer cells. Experimentally, this suggests further sensibly chosen targets to better narrow down the amount of potential causes for the witnessed scenarios. (i) A reasonable extension of metabolic observables can improve the analyses of metabolic fluxes. We recommend more metabolic targets that are situated around complex branching points of central carbon metabolism, such as glucose 6-phosphate (branching to pentose phosphate pathway) and pyruvic acid (with routes to e.g. lactic acid, alanine, acetate, acetaldehyde, acetyl-CoA). (ii) Also, experimental interaction studies of the KRAS and BRAF proteins with the aforementioned enzymes can be of interest, to test for direct interactions with LDH, 2-oxoglutarate:ferredoxin oxidoreductase, or the

pyruvic acid:ferredoxin oxidoreductase. Lastly, (iii) phosphoproteomic targets downstream of the Protein Kinase B pathway can narrow down the search for potential candidates that cause the witnessed metabolic deregulations.

As a technical consideration, the complexity of the regulatory system and the heterogeneity of the experimental data requires a more precise modelling approach than a classical metabolic flux analysis. We have to aim for a more advanced modelling approach to account for potential candidates that cannot be uncovered by flux analyses: The influence of the intricately maintained NADP:NADPH balance of the mitochondrion, for instance, can be a key issue when analysing the displayed scenarios. It is significant, since (i) the synthesis of lactic acid by lactate dehydrogenase is coupled to the conversion of NADPH to NADP [40]. Also, (ii) $\alpha$-ketoglutaric acid is reduced to isocitric acid under the oxidation of NADPH to NADP, which is in this case the initial step of the reductive TCA cycle. The sensitive balance of these involved co-factors demands for a more sophisticated mathematical modelling approach than a classical metabolic flux analysis. A recommendable choice can be a model of ordinary differential equations (ODEs) that allows a dynamic continuous simulation of the system and displays the concentration changes over time. Such a model can consider the balance of co-factors as well as the thermodynamic feasibility of reactions. The benefits of this approach have to be earned by the provision with more experimental data and refined parameter estimation and data augmentation techniques. However, these methods are available [41, 42] and experimental data can be provided as soon as knowledge gaps and suitable targets are identified. With this work, we have aimed at providing suggestions for these gaps and targets at numerous points of cancer cell metabolism.

# References

[1] Arvelo F, Sojo F, Cotte C (2015) Biology of colorectal cancer. ecancermedicalscience 9.

[2] Ying H, Kimmelman AC, Lyssiotis CA, Hua S, Chu GC, et al. (2012) Oncogenic kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. Cell 149: 656–670.

[3] Jaitin D, Sayles L, Goliazova T, Denko N, Sweet-Cordero A (2012) Oncogenic kras inhibits mitochondrial metabolism by regulating the pyruvate dehydrogenase complex under conditions of nutrient stress. Cancer Research 72: 1000.

[4] Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. cell 144: 646–674.

[5] Ward PS, Thompson CB (2012) Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer cell 21: 297–308.

[6] Mor A, Aizman E, George J, Kloog Y (2011) Ras inhibition induces insulin sensitivity and glucose uptake. PloS one 6: e21712.

[7] Bunz F, Papadopoulos N (2012) Ras/raf and their influence in glycolysis in colon cancer. In: Energy Balance and Gastrointestinal Cancer, Springer. pp. 131–139.

[8] Quijano C, Cao L, Fergusson MM, Romero H, Liu J, et al. (2012) Oncogene-induced senescence results in marked metabolic and bioenergetic alterations. Cell Cycle 11: 1383–1392.

[9] Kamphorst JJ, Fan J, Lu W, White E, Rabinowitz JD (2011) Liquid chromatography–high resolution mass spectrometry analysis of fatty acid metabolism. Analytical chemistry 83: 9114–9122.

[10] Hidalgo IJ, Raub TJ, Borchardt RT (1989) Characterization of the human colon carcinoma cell line (caco-2) as a model system for intestinal epithelial permeability. Gastroenterology : 736–49.

[11] Pietzke M, Zasada C, Mudrich S, Kempa S (2014) Decoding the dynamics of cellular metabolism and the action of 3-bromopyruvate and 2-deoxyglucose using pulsed stable isotope-resolved metabolomics. Cancer & metabolism 2: 1–11.

[12] Zamboni N, Fendt SM, Rühl M, Sauer U (2009) 13c-based metabolic flux analysis. Nature protocols 4: 878–892.

[13] Young JD, Walther JL, Antoniewicz MR, Yoo H, Stephanopoulos G (2008) An elementary metabolite unit (emu) based method of isotopically nonstationary flux analysis. Biotechnology and bioengineering 99: 686–699.

[14] Kiefer P, Heinzle E, Zelder O, Wittmann C (2004) Comparative metabolic flux analysis of lysine-producing corynebacterium glutamicum cultured on glucose or fructose. Applied and Environmental Microbiology 70: 229–239.

[15] Klapa MI, Aon JC, Stephanopoulos G (2003) Systematic quantification of complex metabolic flux networks using stable isotopes and mass spectrometry. European Journal of Biochemistry 270: 3525–3542.

[16] Van Winden WA, Wittmann C, Heinzle E, Heijnen JJ (2002) Correcting mass isotopomer distributions for naturally occurring isotopes. Biotechnology and bioengineering 80: 477–479.

[17] Fritsche-Guenther R, Witzel F, Sieber A, Herr R, Schmidt N, et al. (2011) Strong negative feedback from erk to raf confers robustness to mapk signalling. Molecular systems biology 7: 489.

[18] Möller Y, Siegemund M, Beyes S, Herr R, Lecis D, et al. (2014) Egfr-targeted trail and a smac mimetic synergize to overcome apoptosis resistance in kras mutant colorectal cancer cells. Plos One .

[19] Herr R, Wöhrle FU, Danke C, Berens C, Brummer T (2011) A novel mcf-10a line allowing conditional oncogene expression in 3d culture. Cell Communication and Signaling 9: 17.

[20] Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, et al. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Molecular & cellular proteomics 4: 1487–1502.

[21] Wiechert W (2001) 13 c metabolic flux analysis. Metabolic engineering 3: 195–206.

[22] Möllney M, Wiechert W, Kownatzki D, de Graaf AA (1999) Bidirectional reaction steps in metabolic networks: Iv. optimal design of isotopomer labeling experiments. Biotechnology and Bioengineering 66: 86–103.

[23] Sauer U (2006) Metabolic networks in motion: 13c-based flux analysis. Molecular systems biology 2: 62.

[24] MATLAB (2010) version 7.10.0 (R2010a). Natick, Massachusetts: The MathWorks Inc.

[25] Helmert FR (1876) Über die wahrscheinlichkeit der potenzsummen der beobachtungsfehler. Z Math u Phys 21: 192–218.

[26] Riedwyl H (1967) Goodness of fit. Journal of the American Statistical Association 62: 390–398.

[27] Young JD (2014) Inca: a computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics 30: 1333–1335.

[28] Warburg O, et al. (1956) On the origin of cancer cells. Science 123: 309–314.

[29] Tessem MB, Swanson MG, Keshari KR, Albers MJ, Joun D, et al. (2008) Evaluation of lactate and alanine as metabolic biomarkers of prostate cancer using 1h hr-mas spectroscopy of biopsy tissues. Magnetic Resonance in Medicine 60: 510–516.

[30] Cornel E, Smits G, Oosterhof G, Karthaus H, Deburyne F, et al. (1993) Characterization of human prostate cancer, benign prostatic hyperplasia and normal prostate by in vitro 1h and 31p magnetic resonance spectroscopy. The Journal of urology 150: 2019–2024.

[31] Swanson MG, Zektzer AS, Tabatabai ZL, Simko J, Jarso S, et al. (2006) Quantitative analysis of prostate metabolites using 1h hr-mas spectroscopy. Magnetic Resonance in Medicine 55: 1257–1264.

[32] Sabharwal SS, Schumacker PT (2014) Mitochondrial ros in cancer: initiators, amplifiers or an achilles' heel? Nature Reviews Cancer 14: 709–721.

[33] King A, Selak M, Gottlieb E (2006) Succinate dehydrogenase and fumarate hydratase: linking mitochondrial dysfunction and cancer. Oncogene 25: 4675–4682.

[34] Gaglio D, Metallo CM, Gameiro PA, Hiller K, Danna LS, et al. (2011) Oncogenic k-ras decouples glucose and glutamine metabolism to support cancer cell growth. Molecular systems biology 7: 523.

[35] Yun J, Rago C, Cheong I, Pagliarini R, Angenendt P, et al. (2009) Glucose deprivation contributes to the development of kras pathway mutations in tumor cells. Science 325: 1555–1559.

[36] Mazurek S (2011) Pyruvate kinase type m2: a key regulator of the metabolic budget system in tumor cells. The international journal of biochemistry & cell biology 43: 969–980.

[37] Zscharnack K, Kessler R, Bleichert F, Warnke J, Eschrich K (2009) The pfkfb3 splice variant ubi2k4 is downregulated in high-grade astrocytomas and impedes the growth of u87 glioblastoma cells. Neuropathology and applied neurobiology 35: 566–578.

[38] McCleland ML, Adler AS, Deming L, Cosino E, Lee L, et al. (2013) Lactate dehydrogenase b is required for the growth of kras-dependent lung adenocarcinomas. Clinical Cancer Research 19: 773–784.

[39] Huang C, Sheng S, Li R, Sun X, Liu J, et al. (2015) Lactate promotes resistance to glucose starvation via upregulation of bcl-2 mediated by mtor activation. Oncology reports 33: 875–884.

[40] Holmes RS, Goldberg E (2009) Computational analyses of mammalian lactate dehydrogenases: human, mouse, opossum and platypus ldhs. Computational biology and chemistry 33: 379–385.

[41] Lubitz T, Schulz M, Klipp E, Liebermeister W (2010) Parameter balancing in kinetic models of cell metabolism. The Journal of Physical Chemistry B 114: 16298–16303.

[42] Stanford NJ, Lubitz T, Smallbone K, Klipp E, Mendes P, et al. (2013) Systematic construction of kinetic models from genome-scale metabolic networks. Plos One .

## 2.4 SBtab: An SBML-interconvertible table format for data exchange in Systems Biology

During the time of my doctoral studies, I have learned about the importance of workflows and their reproducibility. A major prerequisite to a workflow is an accurate description of the interface formats that are employed for data exchange between the workflow steps. Standard formats are a useful approach for a seamless data exchange; they allow for the combination of software tools and can be processed automatically. Chapter 1.2 offers a short summary on the merits of and requirements for standard formats. However, the existing Systems Biology standard formats each have their shortcomings, like limitations to specific kinds of data, some lack human-readability, others only target very distinct user groups. Furthermore, some can only be used with technical knowledge or software expertise.

We have developed a new interface format - SBtab - which supports the exchange and automated processing of data tables. In general, SBtab files are data tables comprising diverse kinds of information (e.g. lists of reactions, compounds, parameters, etc.), which have to adhere to certain conventions and syntax rules. They can be created with any text editor or spreadsheet tool (like MS Excel) and examples for different predefined data types can be seen in Figure 2.1.

The effort of a new proposal for a standardised data format would be futile, if it could not be processed automatically by software tools. Thus, we are providing a set of versatile tools that facilitate the usage and improve the applicability of SBtab:

**Online Validator** To ensure that an SBtab file follows all defined conventions, users can upload the file to our website `www.sbtab.net`. Here, SBtab files can be displayed and automatically validated to test their syntactic correctness.

**SBML Converter** On the same website, SBtab files with model structure data can be converted to SBML models and *vice versa*.

**SQLite interface** SBtab files can be used for the import and export of data into and from an SQLite database. The database can be queried *via* standard SQL queries, which are supported by most programming languages (this tool is a courtesy of Dr. Elad Noor, ETH Zürich).

**xlSBtab Excel Add-in** The SBtab validator and SBML converter can also be used from within Microsoft Excel *via* our Add-in called "xlSBtab" (this tool is a courtesy of Dr. Frank T. Bergmann, COS Heidelberg).

**Python scripts** Most SBtab tools are also available as Python scripts with easy interfaces. This enables programmers to embed SBtab objects and corresponding functions into their own software tools or workflows.

A

| !!SBtab | TableName='Reaction Example' | TableType='Reaction' | | | | | |
|---|---|---|---|---|---|---|---|
| !Reaction | !SumFormula | !Identifiers:kegg.reaction | !Gene:Symbol | !Location | !Modifier | !IsReversible | !Pathway |
| R1 | ATP + F6P <=> ADP + F16BP | R00658 | pfk | cytosol | ATP | False | glycolysis |
| R2 | F16BP + H20 <=> F6P + Pi | R01015 | fbp | cytosol | | False | glycolysis |

B

| !!SBtab | TableName='Kinetic Parameters' | TableType='Quantity' | | | | | |
|---|---|---|---|---|---|---|---|
| !Quantity | !QuantityType | !Reaction:Identifiers:kegg.reaction | !Compound:Identifiers:kegg.compound | !Value | !Unit | !Temperature | !pH |
| keq_R1 | equilibrium constant | R01061 | | 0.156 | dimensi-onless | 300 | 7 |
| kmc_R1_C1 | Michaelis constant | R01061 | C00003 | 0.96 | mM | 290 | 7 |
| kic_R1_C1 | inhibition constant | R01070 | C00111 | 0.13 | mM | 300 | 7.1 |
| con_C1 | concentration | | C00118 | 0.203 | mM | 295 | 7 |
| ... | ... | ... | ... | ... | ... | ... | .... |

C

| !!SBtab | TableType='QuantityMatrix' | TableName='Metabolomics Data' | UniqueKey='False' |
|---|---|---|---|
| !TimePoint | !Time | >Measurement:Glucose | >Measurement:Fructose |
| T0 | 0.0 | 1.1 | 1.4 |
| T1 | 0.5 | 1.2 | 0.9 |

| !!SBtab | TableType=Quantity' | TableName='Measurement' | UniqueKey='False' |
|---|---|---|---|
| !Compound | !Identifiers:obo.chebi | !QuantityType | !Unit |
| Glucose | CHEBI:17234 | concentration | mM |
| Fructose | CHEBI:15824 | concentration | mM |

**Figure 2.1: SBtab example files.** In all SBtab files, the first table row is preceeded with two exclamation marks and declares table attributes, like the type of included data. The second row holds controlled column headers, each preceded with a single exclamation mark. Furthermore, all depicted examples only show a small fraction of possible predefined columns. (A) The SBtab type *Reaction* enlists information on biochemical reactions; (B) A *Quantity* SBtab comprises kinetic parameter information; (C) The SBtab types *QuantityMatrix* and *Quantity* can reference each other to represent information on experimental time point data in a combined effort. The referencing of related SBtab files from within another file is executed by the symbol ">" in the column header, followed by the referenced SBtab table name and the comprised entry ID (here, the column ">Measurement:Glucose" refers to the entry "Glucose" from the SBtab file "Measurement"). The references are also visualised by red arrows for the sake of clarity.

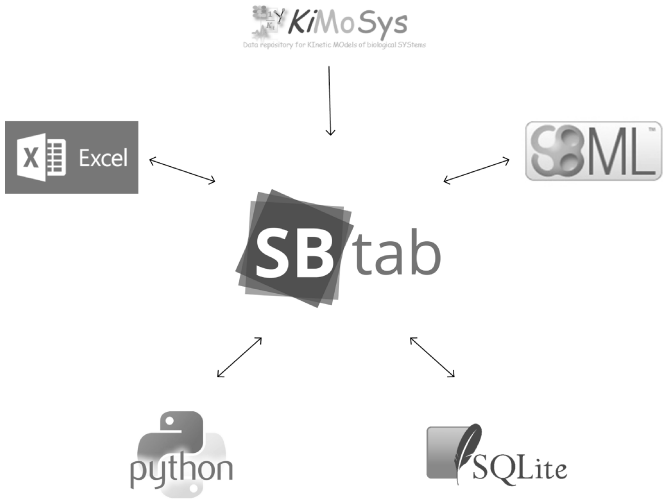



**Figure 2.2: The interfaces of SBtab.** SBtab has interfaces to an illustrious set of tools; this includes the conversion from and to SBML, an add-in for MS Excel, an SQLite database interface, Python programming scripts, and data export from the KiMoSys database[46].

These tools allow the integration of SBtab for numerous applications in Systems Biology workflows: Model creation, model manipulation, incorporation of data into models or databases, and more (see Figure 2.2 for an overview). I have successfully applied several of these approaches during my projects and am still working on the further development of the format. A full specification including predefined table types, structures, and conventions is provided in the corresponding SBtab specification[45], and can be found in Appendix 4.1 of this thesis. It also features guidance on how to customise the SBtab format to individual data types, which are not part of the predefined SBtab table types. I have applied SBtab for numerous efforts of my projects: The pathway reconstruction of Chapter 2.1.1 was formulated in the rxncon format, which is convertable to and from a customised SBtab table type *RxnconTable*. For Chapter 2.2.2, the kinetic data incorporation and parameter balancing were performed with the SBtab table type *QuantityType*. And finally, for Chapter 2.3 I have created a metabolic SBML model *via* the SBtab table types *Reaction*, *Compound*, and *Compartment*.

# SBtab: a flexible table format for data exchange in systems biology

**Authors:** Timo Lubitz, Jens Hahn, Frank T. Bergmann, Elad Noor, Edda Klipp, and Wolfram Liebermeister

**Journal:** This article has been published in Bioinformatics, 2016

**Accessiblity:** It can be accessed online via doi:10.1093/bioinformatics/btw179

## Abstract

**Summary:** SBtab is a table-based data format for Systems Biology, designed to support automated data integration and model building. It uses the structure of spreadsheets and defines conventions for table structure, controlled vocabularies and semantic annotations. The format comes with pre- defined table types for experimental data and SBML-compliant model structures and can easily be customized to cover new types of data.

**Availability and Implementation:** SBtab documents can be created and edited with any text editor or spreadsheet tool. The website www.sbtab.net provides online tools for syntax validation and conversion to SBML and HTML, as well as software for using SBtab in MS Excel, MATLAB and R. The stand-alone Python code contains functions for file parsing, validation, conversion to SBML and HTML and an interface to SQLite databases, to be integrated into Systems Biology workflows. A detailed specification of SBtab, including examples and descriptions of table types and available tools, can be found at www.sbtab.net.

**Contact:** wolfram.liebermeister@gmail.com

# 3 Conclusion

With each chapter, this thesis has subsequently lead through diverse regular subsystems of the cell. An extracellular signal has entered the cell and was transduced over a signalling cascade down to gene regulation in Chapter 2.1. In the following chapter 2.2, the changed gene expression patterns affected metabolic functionality by the alteration of the cells' protein composition. Finally, Chapter 2.3 approached a more holistic view of the cellular system with the focus on oncogene-induced metabolic reprogramming in cancer cells. Figure 3.1 serves as a roadmap for the interconnection of these different subsystems. It remains to be answered, how the proposed modelling workflows contribute to the technical standards of Systems Biology and how the projects' results expand our understanding of the targetted biological systems.

**Exploring signalling pathways** The mathematical modelling of signalling pathways (see left side of Figure 3.1) is an intricate field of research. The interactions of cascade components are hard to measure and underlie highly complex and diverse mechanisms (changes in molecule structure, posttranslational modifications, complex formations, and more). Although the amount of experimental data for signalling pathways increases rapidly, data often is not sufficient for a detailled time-resolved modelling of complete signal transductions from input signal to gene expression[47]. The most common modelling efforts of signalling pathways comprise ODE models of comparably small-scaled systems[48,49,50,51] and also serve network-based methods for larger models[52,53]. Both approaches are justified for different types of analysis, an aspect which I have illustrated within this thesis.

The large-scale modelling of signalling pathways is a growing field in Systems

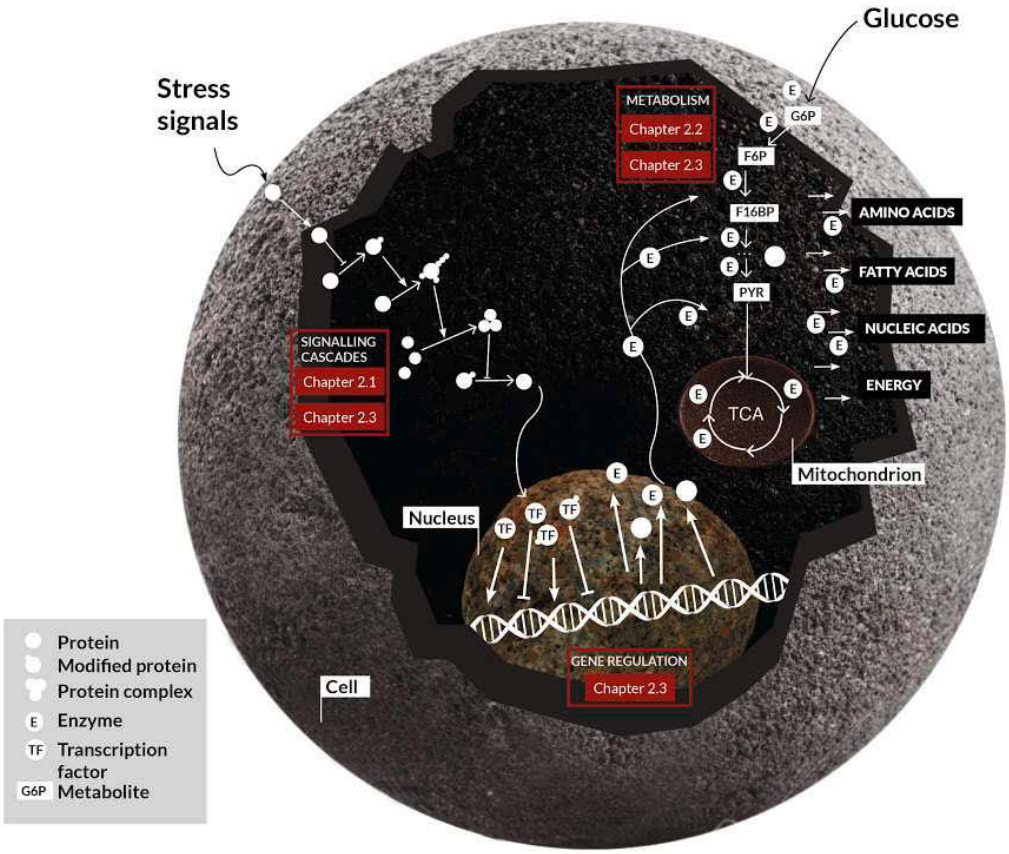



**Figure 3.1: Regulatory layers of the cell. Left**: Extracellular signals are received by the cell and transmitted *via* signalling cascades. The information is passed down to the cells' nucleus by subsequent protein modifications and the formation or breakup of protein complexes. In the nucleus, the activity of transcription factors is altered. **Bottom**: In response to transcription factor activity changes, genes are activated or inhibited. This changes gene expression levels and thus the protein composition of the cell. **Right**: The proteins are governing metabolic functionality. Depending on their expression levels, metabolism converts nutrients to energy and cellular building blocks.

Biology. Samaga et al. present a thorough review on the currently employed formalisms, comprising interaction graphs, Boolean models, and logic-based ODEs[54]. (i) Interaction graphs are suitable for a crude and first representation of the system and they offer simple means for the analysis of the components' dependencies. This, however, limits their applicability: Realistic simulations are not feasible and this eliminates the option of employing them for hypotheses testing. (ii) Boolean models improve this analysis by adding logical combinations of edges and thus modelling an important regulatory layer[29,55]. (iii) Logic-based ODEs can be fitted to experimental data, but they do not add further information to the Boolean logic implementation; they still must be considered phenomenological models[56]. The dependence on experimental data makes logic-based ODEs an interesting option for the future, but it cannot be the formalism for today. In light of these three approaches, we have decided to employ the Boolean solution for the investigation of the Snf1 pathway.

Our introduced approach for constructing large-scale signalling network reconstructions and validating their accuracy and connectivity with Boolean modelling (see Chapter 2.1.1) is novel, flexible, and versatile. It excels from previous approaches by several features: The initial network reconstruction is based on a manual curation of all publicly available Snf1 literature; the employed rxncon framework[57] offers a sophisticated constraint-based modelling approach for signalling networks; the iterative curation process of the Boolean model exposes knowledge gaps of the pathway on the basis of verified input/output relationships of the network; the final model can be exported into a rule-based representation, which builds a bridge from Boolean to dynamic modelling. A bridge that is urgently needed, since Boolean modelling may only be an interim solution for more detailled and continuous modelling efforts. The rule-based model employs 176 distinct parameters, which shows where the next required steps in the modelling of signalling pathways lie. More experimental data will give better insight into the complex and diverse reaction cascades of these pathways, which will eventually allow us to model them dynamically. Until then, reconstructive workflows like the one presented here are the main tool for the exploration of pathways as a whole and the direction of future experiments.

The application of the aforementioned workflow on the Snf1 pathway is the first large-scale modelling effort of this otherwise well-analysed pathway. One of the benefits of the workflow is the exposition of knowledge gaps, and the main gap in the Snf1 pathway is the mechanistic activation of Snf1[58]. For the analysis of this crucial point, I have applied ODEs for the modelling of different hypothetic candidates of Snf1 regulation (see Chapter 2.1.2). This process of model discrimination on the grounds of experimental data is not new[59,60]; but it is a powerful mathematical approach for minimising the distance between computational simulation and reality[61]. Furthermore, there is no mathematical modelling effort for this major part of the Snf1 pathway available so far. And the results of the project still withstand: Our suggestion of a second glucose-regulated step between the activation of Snf1 and

the dephosphorylation of Mig1 has found several candidates in the recent years. The SUMOylation and ubiquitylation of Snf1 by Mms21 and Slx5, respectively[62], as well as the binding of ADP to prevent Snf1 dephosphorylation[63]. The latter is a coherence to the mammalian homologue of Snf1, AMPK, which is regulated by intracellular levels of AMP. But although our findings have contributed to better our understanding of the mechanism of Snf1 activation, we are still on the search for more details of Snf1 regulation. This is a comprehensible effort, since AMPK, the mammalian homologue of Snf1, is an important player in cellular stresses and diseases. Mutations leading to a deregulation of AMPK are the cause of many human lifestyle diseases, such as obesity, type 2 diabetes, cancer, and heart diseases[64]. Thus, Snf1/AMPK is a significant target for regulating drugs and treating related diseases. We need to understand the regulation of Snf1/AMPK activation in mechanistic detail in order to find starting points for drug development. This research is further complicated by the fact that Snf1 is not only activated by glucose deprivation, but also by a various line-up of other stresses, including salt stress, nitrogen starvation, ultraviolet radiation, alkaline pH, and others. The presented results are one more piece for solving the puzzle of Snf1 pathway regulation.

**The intricate enzymatic regulation of metabolic models** Just like the mathematical modelling of signalling pathways, models of cellular metabolism (see right hand side of Figure 3.1) differ in size, scope, and formalism[65]. The most common approaches are ODE models for networks of a smaller scale and stoichioimetric models for genome-scale implementations. While the beginning of the century has brought several small and middle sized glycolysis and central carbon metabolism models to light[66,67,68], metabolomic and genomic data accumulated fast and made way to first stoichioimetric genome-scale reconstructions of cell metabolism[69,70,71]. The smaller models provide a detailled, but necessarily limited insight into cell metabolism. Single reactions or chains of subsequent reactions can be modelled fairly reliable, but they are only a glance through a possibly deceptive keyhole. They neglect branches and interactions with other cellular components, the employed parameters may be measured under differing conditions (e.g. pH and temperature), and these circumstances broaden the gap between *in silico* and *in vivo* even further. In contrast, genome-scaled models do not offer a parametrisation at all due to their complexity. They can be employed for approaches like elementary mode or flux balance analysis, which are efficient methods; but they can only be a weaker alternative to dynamic continuous modelling. In between the small and genome-scale models are the ever-growing and improving approaches for large-scale kinetic modelling of metabolism[72,73,74,26]. All of them, however, need to cope with the same obstacles of large-scale metabolic modelling, such as the choice of kinetic rate laws, proper formulations of a biomass reaction, the availability and interpretability of experimental data, or the sensitivity of parameters. Most of these questions are addressed by our review about the evolution of enzyme kinetics (see Chapter 2.2.1) and an extensive review on the topic of large-scale mo-

delling in yeast was introduced by Österlund et al.[65]. So far, these approaches have been hampered by a lack of steady state or unrealistic behaviour against system perturbations.

With our introduced workflow for large-scale kinetic modelling of metabolism (see Chapter 2.2.2) we aim at overcoming the main obstacles of the field and the presented yeast metabolism model shows significant improvements to previously published models of the same size scale. The application of parameter balancing ensures thermodynamic feasibility and augments the fragmentary kinetic parameter set by employing parameter dependencies within a Bayesian framework. Furthermore, differences in measurement conditions of the parameters can be adjusted. Our model reaches a steady state, which is stable against perturbations of glucose concentration. Furthermore, the performed metabolic control analysis revealed a realistic model behaviour in response to changes in enzyme concentration. All these features have been achieved with experimental and kinetic data from the literature or approximative values, which stands for the prowess of the workflow and gives an estimation of its capabilities if directed experimental data was applied. It also shows that knowledge gaps can be addressed by approximative approaches and well-grounded assumptions. But despite these improvements, it needs to be duly noted that this is not a finished model. We are rather proposing a powerful workflow for the creation of large-scale metabolic models, which can be extended, iterated, and improved as soon as new experimental data becomes available. This workflow for the creation of kinetic models can make them a powerful tool for investigating changes in enzyme concentration. For instance, in the FBA step of the workflow, fluxes may be constrained to 0, representing gene knockouts, i.e. a loss of enzymatic regulation. Until today, further steps have been taken in large-scale modelling by incorporating the ensemble modelling approach[75,76]; focussing on sophisticated parameter estimation techniques to overcome lack of kinetic data[77,78]; combining dynamic and genome-scale approaches with hybrid techniques[79]; by adding knowledge about the gene regulation module[80]; and finally, of course, by the incorporation of more experimental data[81]. A recent review summarises these latest achievements in large-scale kinetic modelling of cell metabolism in detail[82].

It is important that metabolic models show realistic behaviours in order to become meaningful tools in cellular research. Their simulation and application can direct future experiments to gaps in our knowledge of the system. And understanding metabolism as a whole opens possibilities: Metabolic engineering, for instance, supports the production of drugs (like insulin and antibiotics) or of required industrial chemicals (like shikimic acid)[83] *via* sophisticated strain design techniques[84]. But naturally, another reason for our longing to promote metabolic research can be directly adapted from what has been said about signalling pathways: It is once more a potential starting point for drug therapy. Metabolism, as a cellular subsystem, is different from signalling pathways, but the scope of this thesis should have shed light on the direct connection of the two. One cannot understand one without the other, and if we want to thoroughly direct research towards

the starting points of drug development, we need to understand metabolism just as well as cell signalling. The sole focus on metabolism can show us how enzymes control metabolic functionality, why certain nutrients cannot be metabolised, or which metabolic reactions are responsible for the accumulation of metabolic by-products, for example. But the background of this "output" behaviour can only be understood by taking a step back and regarding the larger, more holistic picture of the cell.

**Signalling proteins, gene regulation, and metabolism in a joint modelling approach** Cancer cell metabolism urgently requires a holistic analysis, which comprises aspects from signalling pathways as well as oncogene regulation and metabolism itself (see Figure 3.1). The behaviour of cancer cells is differing strongly to that of normal cells. Their growth factor responses are constantly active, independent of growth factor availability. They prefer aerobic glycolysis irregardless of oxygen levels. Nutrition uptake is upregulated, which is compensated by the accumulation and fermentation of lactic acid[85]. These are only some major differences, which are complemented by many other metabolic deregulations. They are mainly put into effect by oncogenes, which frequently encode proteins from signalling pathways. But it cannot be inferred that signalling pathways are solely responsible for this metabolic reprogramming of cancer cells. Instead, many of the oncogene encoded proteins affect metabolic components in a direct manner instead of altering the course of signalling cascades[86,87]. These coherences make it inevitable to regard cancer cell metabolism in context with a proper analysis of oncogenes and the signalling status of the cell.

To date, there are many different advances in the modelling of cancer-specific systems. Like in normal cells, ODE models are a convenient tool for detailled signalling[88,89], metabolic[90], and growth/population models of a small scale[91,92]. These models are often focussed on the differences between normal and cancerous cells, as well as the introduction of external factors (such as estrogen[93]). Their need for system simplification is bigger than for comparable models of normal cells, which is due to fewer experimental data and thus fewer available knowledge. Furthermore, the specificity of different cancers narrows down the field of available knowledge even more, so that most cancer modelling approaches are in direct connection to specific experimental setups[94,95] (while modellers of healthy cells have the initial opportunity to rely on exhaustive literature searches to collect metabolite concentrations and kinetic parameters[96]). Due to these circumstances, metabolic flux analysis (MFA) has become a convenient tool for reliably investigating details of cancer cell metabolism[97], since they do not require kinetic detail, but still yield realistic impressions of metabolic flux. MFA have been applied in combination with latest $^{13}C$ isotopic tracing experiments[98,99].

The final project of this thesis excels by employing a more holistic view than the previous approaches. I perform non-stationary metabolic flux analyses for colorectal cancer cell lines exhibiting mutations of two common oncogenes (KRAS

and BRAF). Furthermore, the cells are incorporated with $^{13}C$ glucose and glutamine, which enables us to examine details of glycolysis as well as glutaminolysis in the cells. The results are augmented by proteomic and phosphoproteomic data, which have been taken simultaneously to the metabolomic data. The results clearly link the KRAS protein to the accumulation of lactic acid in cancer cells, a key feature of the Warburg effect. Furthermore, the carbon routing through the reductive TCA cycle could be connected to the KRAS protein as well as BRAF. These proteins are encoded by common oncogenes that are on the one hand important players in signalling pathways and, on the other hand, known to have direct influences on cancer cell metabolism[87]. Finally, we are analysing the regulatory system on grounds of the provided data and propose suitable experimental targets as well as appropriate mathematical modelling techniques to improve our results so far. With the presented results we contribute important novelties about the details of metabolic reprogramming in cancer cells.

The proposed results are of large interest and stand for the latest hallmark in cancer research: metabolic reprogramming[87]. While more and more details on varying enzyme concentrations in cancer cells are revealed, the causes for these deviations from the norm are largely unknown and mostly attributed to oncogenes. The need for the exploration of these details is evident: Cancer is one of the most wide-spread diseases and takes second place on the list of most frequent death causes after heart disease[100]. Its nature is so heterogenous among different cancer types and so much deviating from that of normal cells that cancer research still is a long way from personalised treatment possibilities. Systems Biology, however, has introduced new and favourable mathematical approaches, which, in combination with new experimental high-throughput techniques, can accelerate the speed of cancer research significantly[95,101].

**Concluding words and outlook** The presented projects of my thesis comprise (i) novel workflows for creating large-scale signalling pathway reconstructions and kinetic metabolic models; (ii) new biological insight into the regulatory principles of yeast and human cells; and finally (iii) a proposal for a standardised table format, which facilitates the exchange and automated usage of data in Systems Biology.

The knowledge database which we have acquired in an extensive literature curation on the Snf1 pathway in yeast is soon going to be extended for the more recent developments in this field of research. The database will then be the foundation of a comprehensive review on the subject, which is a valuable compendium of manually curated knowledge on one of the most prominent pathways in yeast.

Our ongoing project on cancer cell metabolism will be refined and extended. From the modelling side, I am currently working on the dynamic ODE model suggested in the corresponding chapter. For this, I am orienting on our proposed metabolic modelling framework of Chapter 2.2.2, where the results of a flux analysis method (in this case the metabolic fluxes) are employed as input for the parameter balancing process. Common modular rate laws are a good approximative choice

for the reaction kinetics, where we can also easily employ the proteomic concentration data as enzymatic prefactors. Furthermore, these choices of rate law and modelling method will provide us valuable insight into the dynamics and regulatory principles of the system. From the experimental side, we have already taken the MFA results as input for what can be proclaimed another round of the iterative cycle of Systems Biology: Conducting new experiments. New measurements of further observables in cancer cell metabolism are currently ongoing and might give us even more insight into the system. And understanding a disease is a major prerequisite of eventually curing it.

Our table format, SBtab, has already been employed by several researchers from different parts of the world. Currently, we are near completion of an adaptation of the SBtab format to suit the specific requirements of the rxncon software framework; tailoring the SBtab format conventions to individual data formats is one of SBtabs inherent features. The technical applicability of SBtab and its practical benefits to science still need to be proven after the manuscript, which is currently under revision at *Bioinformatics Journal*, will be published. We are still working on improving and promoting this ongoing project and consider it a valuable addition to Systems Biology research.

The presented workflows and projects are aimed at improving our computational modelling approaches and thus further our understanding of biological systems. As declared earlier, the benefits and implications of understanding these systems are numerous and significant. And although the road to unravelling the enigmas of life is still long and probably endless, I have contributed my time and work to overcome a few of the required steps on this road.

# References


[1] H. Lodish, *Molecular cell biology*. Macmillan, 2008.

[2] S. Fields and M. Johnston, "Whither model organism research," *Science*, vol. 307, no. 5717, pp. 1885–1886, 2005.

[3] S. B. Hedges, "The origin and evolution of model organisms," *Nature Reviews Genetics*, vol. 3, no. 11, pp. 838–849, 2002.

[4] U. Alon, *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.

[5] B. N. Kholodenko, "Cell-signalling dynamics in time and space," *Nature reviews Molecular cell biology*, vol. 7, no. 3, pp. 165–176, 2006.

[6] K. Takahashi, S. N. V. Arjunan, and M. Tomita, "Space in systems biology of signaling pathways–towards intracellular molecular crowding in silico," *FEBS letters*, vol. 579, no. 8, pp. 1783–1788, 2005.

[7] S. A. Henry, S. D. Kohlwein, and G. M. Carman, "Metabolism and regulation of glycerolipids in the yeast saccharomyces cerevisiae," *Genetics*, vol. 190, no. 2, pp. 317–349, 2012.

[8] J. Berg and J. Tymoczko, "Stryer: Biochemistry," 2002.

[9] F. Crick *et al.*, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[10] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–741, 2003.

[11] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47–C52, 1999.

[12] P. E. Purnick and R. Weiss, "The second wave of synthetic biology: from modules to systems," *Nature reviews Molecular cell biology*, vol. 10, no. 6, pp. 410–422, 2009.

[13] L. Michaelis and M. L. Menten, "Die kinetik der invertinwirkung," *Biochem. z*, vol. 49, no. 333-369, p. 352, 1913.

[14] N. Bohr, "I. on the constitution of atoms and molecules," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 26, no. 151, pp. 1–25, 1913.

[15] J. D. Watson, F. H. Crick, *et al.*, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[16] I. Wilmut, A. Schnieke, J. McWhir, A. Kind, and K. Campbell, "Viable offspring derived from fetal and adult mammalian cells," *Cloning and stem cells*, vol. 9, no. 1, pp. 3–7, 2007.

[17] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[18] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[19] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D344–D350, 2008.

[20] N. Le Novere, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, *et al.*, "Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D689–D691, 2006.

[21] D. Noble, "Cardiac action and pacemaker potentials based on the hodgkin-huxley equations," 1960.

[22] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[23] E. Klipp, W. Liebermeister, A. Helbig, A. Kowald, and J. Schaber, "Systems biology standards—the community speaks," *Nature biotechnology*, vol. 25, no. 4, pp. 390–391, 2007.

[24] F. J. Bruggeman and H. V. Westerhoff, "The nature of systems biology," *TRENDS in Microbiology*, vol. 15, no. 1, pp. 45–50, 2007.

[25] J. Malmström, H. Lee, and R. Aebersold, "Advances in proteomic workflows for systems biology," *Current opinion in biotechnology*, vol. 18, no. 4, pp. 378–384, 2007.

[26] P. Li, J. O. Dada, D. Jameson, I. Spasic, N. Swainston, K. Carroll, W. Dunn, F. Khan, N. Malys, H. L. Messiha, *et al.*, "Systematic integration of experimental data and models in systems biology," *BMC bioinformatics*, vol. 11, no. 1, p. 582, 2010.

[27] B. O. Palsson, *Systems biology*. Cambridge university press, 2015.

[28] G. E. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[29] R.-S. Wang, A. Saadatpour, and R. Albert, "Boolean modeling in systems biology: an overview of methodology and applications," *Physical biology*, vol. 9, no. 5, p. 055001, 2012.

[30] T. Mori, M. Flöttmann, M. Krantz, T. Akutsu, and E. Klipp, "Stochastic simulation of boolean rxncon models: towards quantitative analysis of large signaling networks," *BMC systems biology*, vol. 9, no. 1, p. 45, 2015.

[31] C. W. Gear and L. R. Petzold, "Ode methods for the solution of differential/algebraic systems," *SIAM Journal on Numerical Analysis*, vol. 21, no. 4, pp. 716–728, 1984.

[32] A. Polynikis, S. Hogan, and M. di Bernardo, "Comparing different ode modelling approaches for gene regulatory networks," *Journal of theoretical biology*, vol. 261, no. 4, pp. 511–530, 2009.

[33] D. J. Wilkinson, *Stochastic modelling for systems biology*. CRC press, 2011.

[34] J. S. Edwards, M. Covert, and B. Palsson, "Metabolic modelling of microbes: the flux-balance approach," *Environmental microbiology*, vol. 4, no. 3, pp. 133–140, 2002.

[35] J. Schwender, J. Ohlrogge, and Y. Shachar-Hill, "Understanding flux in plant metabolic networks," *Current opinion in plant biology*, vol. 7, no. 3, pp. 309–317, 2004.

[36] D. Machado, R. S. Costa, M. Rocha, E. C. Ferreira, B. Tidor, and I. Rocha, "Modeling formalisms in systems biology," *AMB Express*, vol. 1, no. 1, pp. 1–14, 2011.

[37] E. Klipp, W. Liebermeister, A. Helbig, A. Kowald, and J. Schaber, "Standards in computational systems biology," 2007.

[38] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, *et al.*, "The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[39] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[40] S.-A. Sansone, P. Rocca-Serra, M. Brandizi, A. Brazma, D. Field, J. Fostel, A. G. Garrow, J. Gilbert, F. Goodsaid, N. Hardy, *et al.*, "The first rsbi (isa-tab) workshop:"can a simple format work for complex studies?"," *OMICS A Journal of Integrative Biology*, vol. 12, no. 2, pp. 143–149, 2008.

[41] N. Le Novere, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, *et al.*, "The systems biology graphical notation," *Nature biotechnology*, vol. 27, no. 8, pp. 735–741, 2009.

[42] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer, "Copasi—a complex pathway simulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, 2006.

[43] A. Brazma, M. Krestyaninova, and U. Sarkans, "Standards for systems biology," *Nature Reviews Genetics*, vol. 7, no. 8, pp. 593–605, 2006.

[44] J. Märquez and R. Serrano, "Multiple transduction pathways regulate the sodium-extrusion gene pmr2/ena1 during salt stress in yeast," *FEBS letters*, vol. 382, no. 1, pp. 89–92, 1996.

[45] W. Liebermeister, T. Lubitz, and J. Hahn, "Sbtab-conventions for structured data tables in systems biology," *arXiv preprint arXiv:1502.01463*, 2015.

[46] R. S. Costa, A. Veríssimo, and S. Vinga, "Kimosys: a web-based repository of experimental data for kinetic models of biological systems," *BMC systems biology*, vol. 8, no. 1, p. 85, 2014.

[47] B. B. Aldridge, J. M. Burke, D. A. Lauffenburger, and P. K. Sorger, "Physicochemical modelling of cell signalling pathways," *Nature cell biology*, vol. 8, no. 11, pp. 1195–1203, 2006.

[48] B. N. Kholodenko, "Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades," *European Journal of Biochemistry*, vol. 267, no. 6, pp. 1583–1588, 2000.

[49] J. Vilar, R. Jansen, and C. Sander, "Signal processing in the tgf-beta superfamily ligand-receptor network," *PLoS Comput Biol*, vol. 2, no. 1, p. e3, 2006.

[50] A. R. Sedaghat, A. Sherman, and M. J. Quon, "A mathematical model of metabolic insulin signaling pathways," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 283, no. 5, pp. E1084–E1101, 2002.

[51] X. Ouyang, X. Huang, X. Jin, Z. Chen, P. Yang, H. Ge, S. Li, and X. W. Deng, "Coordinated photomorphogenic uv-b signaling network captured by mathematical modeling," *Proceedings of the National Academy of Sciences*, vol. 111, no. 31, pp. 11539–11544, 2014.

[52] G. Liu, A. Marras, and J. Nielsen, "The future of genome-scale modeling of yeast through integration of a transcriptional regulatory network," *Quantitative Biology*, vol. 2, no. 1, pp. 30–46, 2014.

[53] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Random boolean network models and the yeast transcriptional network," *Proceedings of the National Academy of Sciences*, vol. 100, no. 25, pp. 14796–14799, 2003.

[54] R. Samaga and S. Klamt, "Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks," *Cell Commun Signal*, vol. 11, no. 1, p. 43, 2013.

[55] M. K. Morris, J. Saez-Rodriguez, P. K. Sorger, and D. A. Lauffenburger, "Logic-based models for the analysis of cell signaling networks," *Biochemistry*, vol. 49, no. 15, pp. 3216–3224, 2010.

[56] D. M. Wittmann, J. Krumsiek, J. Saez-Rodriguez, D. A. Lauffenburger, S. Klamt, and F. J. Theis, "Transforming boolean models to continuous models: methodology and application to t-cell receptor signaling," *BMC systems biology*, vol. 3, no. 1, p. 98, 2009.

[57] C.-F. Tiger, F. Krause, G. Cedersund, R. Palmér, E. Klipp, S. Hohmann, H. Kitano, and M. Krantz, "A framework for mapping, visualisation and automatic model creation of signal-transduction networks," *Molecular systems biology*, vol. 8, no. 1, 2012.

[58] D. G. Hardie, "Amp-activated/snf1 protein kinases: conserved guardians of cellular energy," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 10, pp. 774–785, 2007.

[59] W. J. Hill, W. G. Hunter, and D. W. Wichern, "A joint design criterion for the dual problem of model discrimination and parameter estimation," *Technometrics*, vol. 10, no. 1, pp. 145–160, 1968.

[60] P. Reilly, "Statistical methods in model discrimination," *The Canadian journal of chemical engineering*, vol. 48, no. 2, pp. 168–173, 1970.

[61] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling, "Ensemble modeling for analysis of cell signaling dynamics," *Nature biotechnology*, vol. 25, no. 9, pp. 1001–1006, 2007.

[62] K. J. Simpson-Lavy and M. Johnston, "Sumoylation regulates the snf1 protein kinase," *Proceedings of the National Academy of Sciences*, vol. 110, no. 43, pp. 17432–17437, 2013.

[63] F. V. Mayer, R. Heath, E. Underwood, M. J. Sanders, D. Carmena, R. R. McCartney, F. C. Leiper, B. Xiao, C. Jing, P. A. Walker, *et al.*, "Adp regulates snf1, the<

i> saccharomyces cerevisiae</i> homolog of amp-activated protein kinase," *Cell metabolism*, vol. 14, no. 5, pp. 707–714, 2011.

[64] D. G. Hardie, F. A. Ross, and S. A. Hawley, "Ampk: a nutrient and energy sensor that maintains energy homeostasis," *Nature reviews Molecular cell biology*, vol. 13, no. 4, pp. 251–262, 2012.

[65] T. Österlund, I. Nookaew, and J. Nielsen, "Fifteen years of large scale metabolic modeling of yeast: developments and impacts," *Biotechnology advances*, vol. 30, no. 5, pp. 979–988, 2012.

[66] B. Teusink, J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. van Dam, H. V. Westerhoff, *et al.*, "Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry," *European Journal of Biochemistry*, vol. 267, no. 17, pp. 5313–5329, 2000.

[67] F. Hynne, S. Danø, and P. G. Sørensen, "Full-scale model of glycolysis in saccharomyces cerevisiae," *Biophysical chemistry*, vol. 94, no. 1, pp. 121–163, 2001.

[68] L. Pritchard and D. B. Kell, "Schemes of flux control in a model of saccharomyces cerevisiae glycolysis," *European journal of biochemistry*, vol. 269, no. 16, pp. 3894–3904, 2002.

[69] B. D. Heavner, K. Smallbone, B. Barker, P. Mendes, and L. P. Walker, "Yeast 5–an expanded reconstruction of the saccharomyces cerevisiae metabolic network," *BMC systems biology*, vol. 6, no. 1, p. 55, 2012.

[70] N. Swainston, K. Smallbone, P. Mendes, D. Kell, and N. Paton, "The subliminal toolbox: automating steps in the reconstruction of metabolic networks," *J Integr Bioinform*, vol. 8, no. 2, p. 186, 2011.

[71] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, *et al.*, "A community-driven global reconstruction of human metabolism," *Nature biotechnology*, vol. 31, no. 5, pp. 419–425, 2013.

[72] R. Steuer, T. Gross, J. Selbig, and B. Blasius, "Structural kinetic modeling of metabolic networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 11868–11873, 2006.

[73] K. Smallbone and N. J. Stanford, "Kinetic modeling of metabolic pathways: Application to serine biosynthesis," in *Systems Metabolic Engineering*, pp. 113–121, Springer, 2013.

[74] K. Smallbone, E. Simeonidis, N. Swainston, and P. Mendes, "Towards a genome-scale kinetic model of cellular metabolism," *BMC Systems Biology*, vol. 4, no. 1, p. 6, 2010.

[75] Y. Tan and J. C. Liao, "Metabolic ensemble modeling for strain engineers," *Biotechnology journal*, vol. 7, no. 3, pp. 343–353, 2012.

[76] A. Khodayari, A. R. Zomorrodi, J. C. Liao, and C. D. Maranas, "A kinetic model of escherichia coli core metabolism satisfying multiple sets of mutant flux data," *Metabolic engineering*, vol. 25, pp. 50–62, 2014.

[77] A. F. Villaverde, D. Henriques, K. Smallbone, S. Bongard, J. Schmid, D. Cicin-Sain, A. Crombach, J. Saez-Rodriguez, K. Mauch, E. Balsa-Canto, *et al.*, "Biopredyn-bench: a suite of benchmark problems for dynamic modelling in systems biology," *BMC systems biology*, vol. 9, no. 1, p. 8, 2015.

[78] A. Gábor and J. R. Banga, "Robust and efficient parameter estimation in dynamic models of biological systems," *BMC systems biology*, vol. 9, no. 1, p. 74, 2015.

[79] K. Tummler, C. Kühn, and E. Klipp, "Dynamic metabolic models in context: biomass backtracking," *Integrative Biology*, vol. 7, no. 8, pp. 940–951, 2015.

[80] K. Smallbone and P. Millard, "Multi-scale modelling of e. coli metabolism," tech. rep., PeerJ PrePrints, 2015.

[81] A. Bordbar, D. McCloskey, D. C. Zielinski, N. Sonnenschein, N. Jamshidi, and B. O. Palsson, "Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics," *Cell Systems*, vol. 1, no. 4, pp. 283–292, 2015.

[82] S. Srinivasan, W. R. Cluett, and R. Mahadevan, "Constructing kinetic models of metabolism at genome-scales: A review," *Biotechnology journal*, vol. 10, no. 9, pp. 1345–1359, 2015.

[83] Y. Matsuoka and K. Shimizu, "Current status and future perspectives of kinetic modeling for the cell metabolism with incorporation of the metabolic regulation mechanism," *Bioresources and Bioprocessing*, vol. 2, no. 1, pp. 1–19, 2015.

[84] N. J. Stanford, P. Millard, and N. Swainston, "Robokod: microbial strain design for (over) production of target compounds," *Frontiers in cell and developmental biology*, vol. 3, 2015.

[85] M. Lopez-Lazaro, "The warburg effect: why and how do cancer cells activate glycolysis in the presence of oxygen?," *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, vol. 8, no. 3, pp. 305–312, 2008.

[86] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[87] P. S. Ward and C. B. Thompson, "Metabolic reprogramming: a cancer hallmark even warburg did not anticipate," *Cancer cell*, vol. 21, no. 3, pp. 297–308, 2012.

[88] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinò, A. Flacco, and P. Valigi, "Computational model of egfr and igf1r pathways in lung cancer: a systems biology approach for translational oncology," *Biotechnology advances*, vol. 30, no. 1, pp. 142–153, 2012.

[89] C. Terfve, T. Cokelaer, D. Henriques, A. MacNamara, E. Goncalves, M. K. Morris, M. van Iersel, D. A. Lauffenburger, and J. Saez-Rodriguez, "Cellnoptr: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms," *BMC systems biology*, vol. 6, no. 1, p. 133, 2012.

[90] K. Smallbone, R. A. Gatenby, R. J. Gillies, P. K. Maini, and D. J. Gavaghan, "Metabolic changes during carcinogenesis: potential impact on invasiveness," *Journal of theoretical biology*, vol. 244, no. 4, pp. 703–713, 2007.

[91] R. Araujo and D. McElwain, "A history of the study of solid tumour growth: the contribution of mathematical modelling," *Bulletin of mathematical biology*, vol. 66, no. 5, pp. 1039–1091, 2004.

[92] F. Kozusko and M. Bourdeau, "A unified model of sigmoid tumour growth based on cell proliferation and quiescence," *Cell proliferation*, vol. 40, no. 6, pp. 824–834, 2007.

[93] C. Mufudza, W. Sorofa, and E. T. Chiyaka, "Assessing the effects of estrogen on the dynamics of breast cancer," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[94] A. R. Anderson and V. Quaranta, "Integrative mathematical oncology," *Nature Reviews Cancer*, vol. 8, no. 3, pp. 227–234, 2008.

[95] P. K. Kreeger and D. A. Lauffenburger, "Cancer systems biology: a network modeling perspective," *Carcinogenesis*, vol. 31, no. 1, pp. 2–8, 2010.

[96] N. J. Stanford, T. Lubitz, K. Smallbone, E. Klipp, P. Mendes, and W. Liebermeister, "Systematic construction of kinetic models from genome-scale metabolic networks," *PloS one*, vol. 8, no. 11, p. e79195, 2013.

[97] U. Sauer, "Metabolic networks in motion: 13c-based flux analysis," *Molecular systems biology*, vol. 2, no. 1, p. 62, 2006.

[98] C. M. Metallo, J. L. Walther, and G. Stephanopoulos, "Evaluation of 13 c isotopic tracers for metabolic flux analysis in mammalian cells," *Journal of biotechnology*, vol. 144, no. 3, pp. 167–174, 2009.

[99] D. Gaglio, C. M. Metallo, P. A. Gameiro, K. Hiller, L. S. Danna, C. Balestrieri, L. Alberghina, G. Stephanopoulos, and F. Chiaradonna, "Oncogenic k-ras decouples glucose and glutamine metabolism to support cancer cell growth," *Molecular systems biology*, vol. 7, no. 1, p. 523, 2011.

[100] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," CA: *a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.

[101] J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, and J. Lankelma, "Cancer: a systems biology disease," *Biosystems*, vol. 83, no. 2, pp. 81–90, 2006.

# 4 Appendix

## 4.1 SBtab specification

The specification of the SBtab format elucidates the conventions and structures of SBtab files. Furthermore, it enlists all predefined table types for different types of information, introduces available tools, and offers guidance for the customisation of the SBtab format to individual data types. The specification is attached in the most recent version 0.9, and it is publicly available on

`http://arxiv.org/abs/1502.01463`.

arXiv:1502.01463v2 [q-bio.MN] 29 Sep 2015

# Conventions for structured data tables in Systems Biology – SBtab version 0.9

Wolfram Liebermeister[1], Timo Lubitz[2], and Jens Hahn[2]

[1] Institut für Biochemie, Charité - Universitätsmedizin Berlin
[2] Institut für Biophysik, Humboldt-Universität zu Berlin

**Abstract**

Data tables in the form of spreadsheets or delimited text files are the most utilised data format in Systems Biology. However, they are often not sufficiently structured and lack clear naming conventions that would be required for modelling. We propose the SBtab format as an attempt to establish an easy-to-use table format that is both flexible and clearly structured. It comprises defined table types for different kinds of data; syntax rules for usage of names, shortnames, and database identifiers used for annotation; and standardised formulae for reaction stoichiometries. Predefined table types can be used to define biochemical network models and the biochemical constants therein. The user can also define own table types, adjusting SBtab to other types of data. Software code, tools, and further information can be found at www.sbtab.net.

## 1 Introduction

Spreadsheets and delimited text tables are the most utilised data formats in Systems Biology. They are easy to use and can hold various types of data. Tables can not only store omics data, but also metabolic network models described by lists of biochemical reactions. However, when tables are exchanged within scientific collaborations, modellers usually prefer tables that can be processed automatically, and the flexibility of spreadsheets can become a disadvantage. If table structures and nomenclature vary from case to case, parsing becomes laborious and new files require new parsers. Furthermore, different naming conventions – for instance, for biochemical compounds – make it hard to combine data, for instance metabolic network models and omics data produced by different researchers. Therefore, rules for structuring tables and for consistent naming and annotations can make tables much more useful as exchange formats in Systems Biology collaborations and for usage in software tools. The SBtab format comprises a set of conventions for data tables that are supposed to make tables easier and safer to work with. Let us start with a couple of examples. Then we continue with a more formal specification of SBtab version 0.9.

**Example 1: Structure of a metabolic network model** A stoichiometric metabolic model can be defined by a list of biochemical reaction formulae, specifying the substrates, products, and their stoichiometric coefficients. Such reactions can be listed in a single column of a spreadsheet, and additional information may be provided: each reaction can have a number or identifier (defined only within the model) and can be linked to an entry in the database KEGG Reaction [1]. Furthermore, reactions may be catalysed by enzymes, which relates them to certain genes. All information could be stored in the following table:

| Reaction | Sum formula | KEGG ID | Gene symbol |
|---|---|---|---|
| R1 | ATP + F6P <=> ADP + F16P | R00658 | pfk |
| R2 | F16P + H2O <=> F6P + Pi | R01015 | fbp |

where ATP, F6P, ADP, F16P, H2O, and Pi are shortnames for metabolites to be used in the model. Although the information is complete and unambiguous, the parser still has to recognise that the columns Sum formula and KEGG ID contain reaction formulae and identifiers in certain formats. If the column names and the syntax of the reaction formulae vary from table to table (e.g. <-> is used instead of <=>), parsing becomes tedious. In the SBtab format, the table would look a little more complicated, but is easy to parse automatically:

| !!SBtab | TableName='Ex 1 - Reaction' | TableType='Reaction' | |
|---|---|---|---|
| !Reaction | !SumFormula | !Identifiers:kegg.reaction | !Gene:Symbol |
| R1 | ATP + F6P <=> ADP + F16P | R00658 | pfk |
| R2 | F16P + H2O <=> F6P + Pi | R01015 | fbp |

In this table, elements highlighted by colours have special meanings (the colours themselves are just used in this text and are not part of the SBtab format). The SBtab table differs from the original table in several ways: the first line (starting with `!!`) declares that the table is an SBtab table of the type `Reaction` and must therefore satisfy syntax rules for this table type. The following line contains the column headers. They start with the `!` character, emphasising that they were not chosen *ad hoc* by the user, but stem from a controlled vocabulary. The predefined column headers do not contain whitespaces. The header `KEGG ID` has been replaced by the term `!Identifiers:kegg.reaction`. This may look complicated, but it allows parsers to retrieve further data from databases in a stable way[1]. The syntax of the reaction formulae is also uniquely defined. In particular, the shortnames of metabolites must not contain any whitespaces or special characters, which simplifies parsing and makes them suitable as variable names for computer models. The meaning of these shortnames can be defined by providing standardised names or database identifiers in a second table of type `Compound`. The compound shortnames will then serve as keys to rows of this table.

| !!SBtab | TableName='Ex 2 - Compound' | TableType='Compound' |
|---|---|---|
| !Compound | !Name | !Identifiers:kegg.compound |
| F6P | Fructose 6-phosphate | C05345 |
| ATP | ATP | C00002 |
| ADP | ADP | C00008 |
| F16P | Fructose 1,6-bisphosphate | C00354 |
| H2O | Water | C00001 |
| Pi | Inorganic phosphate | C00009 |
| PEP | Phosphoenolpyruvate | C00074 |
| AMP | AMP | C00020 |

Both tables together form an SBtab document describing a model. In practice, they can be stored as separate files, as sheets of a spreadsheet file, or within a single table. The following example contains all necessary information to build a stoichiometric model in the SBML (Systems Biology Markup Language) format [3]:

| !!SBtab | TableName='Ex 3 - Reaction' | TableType='Reaction' | |
|---|---|---|---|
| !Reaction | !SumFormula | !Identifiers:kegg.reaction | !SBML:reaction:id |
| R1 | ATP + F6P <=> ADP + F16P | R00658 | r1 |
| R2 | F16P + H2O <=> F6P + Pi | R01015 | r2 |
| !!SBtab | TableName='Ex 3 - Compound' | TableType='Compound' | |
| !Compound | !Name | !Identifiers:kegg.compound | !SBML:species:id |
| F6P | Fructose 6-phosphate | C05345 | f6p |
| ATP | ATP | C00002 | atp |
| ADP | ADP | C00008 | adp |
| ... | ... | ... | ... |

Here, we have added new identifiers (in the columns `SBML:reaction:id` and `SBML:species:id`) for `Reaction` and `Compound` entries to be used in SBML. Such extra names could be necessary if the original shortnames do not comply with SBML's rules for element identifiers.

**Example 2: Table of kinetic constants** In a second example, we specify numerical parameters, for example kinetic constants and metabolite concentrations that appear in a kinetic model. Each quantity can be related to a compound (e.g. a concentration), to a reaction (e.g. an equilibrium constant), or to several biological elements (e.g. to an enzyme and a compound, in the case of Michaelis-Menten constants). As in the previous example, these elements can be specified by unique identifiers, e.g. KEGG compound or reaction identifiers. Furthermore, each quantity has a value and a physical unit. In the SBtab format, we arrange this information in a table of type `Quantity`. Each row contains all information about one of the quantities:

[1]The expression `kegg.reaction` is defined by the MIRIAM resources and used within SBtab. The URL of the KEGG database, defining the identifiers, may change in the future; however, KEGG's Miriam ID (provided by the the MIRIAM resources web service [2]) is guaranteed to remain stable in time.

| !!SBtab | TableName='Ex 4 - Quantity' | TableType='Quantity' | | | |
|---|---|---|---|---|---|
| !Quantity | !QuantityType | !Reaction:Identifiers:kegg.reaction | !Compound:Identifiers:kegg.compound | !Value | !Unit |
| keq_R1 | equilibrium constant | R01061 | | 0.156 | dimensionless |
| kmc_R1_C1 | Michaelis constant | R01061 | C00003 | 0.96 | mM |
| kic_R1_C1 | inhibition constant | R01070 | C00111 | 0.13 | mM |
| con_C1 | concentration | | C00118 | 0.203 | mM |
| ... | ... | ... | ... | ... | ... |

The first two columns specify a name and a type for each quantity. The quantity types (`substrate catalytic rate constant`, `equilibrium constant` etc.) are not chosen *ad hoc*, but stem from the Systems Biology Ontology (SBO) [4]. This ensures a unique spelling and allows software to retrieve definitions and further information from the SBO web services. The biological elements (in this case, reactions, compounds, or both) are specified in the following two columns by unique identifiers from the KEGG database. Columns with human-readable names, or identifiers from other databases, could be added. Unnecessary fields remain empty. The column name `Value` – like some other mathematical terms – is defined for SBtab (arbitrary values in this example). Unit names are defined as in SBML (see below). If the table is used together with a metabolic model, we can use compound and reaction identifiers from the model instead of the Identifiers.org annotations [5]. In this case, the table would read:

| !!SBtab | TableName='Ex 5 - MyData' | TableType='Quantity' | | | |
|---|---|---|---|---|---|
| !Quantity | !QuantityType | !SBML:reaction:id | !SBML:species:id | !Value | !Unit |
| MyData_1 | equilibrium constant | r1 | | 0.156 | dimensionless |
| MyData_2 | Michaelis constant | r1 | atp | 0.96 | mM |
| MyData_3 | inhibition constant | r1 | atp | 0.13 | mM |
| MyData_4 | concentration | | atp | 1.5 | mM |
| ... | ... | ... | ... | ... | ... |

This table, together with a stoichiometric model and a choice of standardised rate laws (like the modular rate laws [6]) completely defines a kinetic metabolic model.

**Example 3: A table with metabolome data** As a last example, let us consider a table with metabolome time series data. For the sake of simplicity, only two metabolites (rows) and measured samples (columns) are shown:

| !!SBtab | TableType='QuantityMatrix' | TableName='Ex 6 - Metabolomics data' | UniqueKey='False' | |
|---|---|---|---|---|
| !Compound | !Identifiers:obo.chebi | t = 0 s | t = 0.5 s | .. |
| Glucose | CHEBI:17234 | 1.1 | 1.2 | .. |
| Fructose | CHEBI:15824 | 1.4 | 0.9 | .. |
| .. | .. | .. | .. | .. |

Tables of this sort can be also be used for other kinds of omics data. In this example, the headers of data columns (e.g., `t = 0 s`) do not follow a specific syntax and contain relevant information (time point and time unit). We shall see below how such information can be provided in SBtab in a more structured manner.

In the following sections, we introduce the general SBtab rules (specification for SBtab version 0.9), as well as formats and conventions for different types of use (see Section 2). It defines a list of table types (see Section 3) and explains the syntax of reaction formulae in the SBtab format (see Section 2.5). Finally, the specification references the available online tools for the handling of SBtab files (see Section 5) and includes an overview of all available SBtab table types in appendix B. Appendix C lists controlled vocabularies and database resources recommended to be used within SBtab.

# 2 Overview of the SBtab format

## 2.1 Basic conventions

SBtab comprises a list of conventions about the structure, nomenclature, syntax, and annotations in tables describing biochemical network models, kinetic parameters, and dynamic data. It contains

1. General rules for the **structure of tables** and the **syntax** used in table fields.
2. Defined **table types** for different kinds of information, each with possible **columns** with defined names and data types (see Table 1; An overview of all predefined table types and their possible columns is given in the appendix).
3. A **syntax for biochemical element annotations** pointing to databases or ontologies.
4. Rules for usage of **names**, **shortnames**, and **database identifiers** used for annotation.
5. **Naming rules for biochemical quantities** to specify the quantities, physical units, and mathematical terms (like `Mean` for mean values).
6. A syntax for **reaction sum formulae**.
7. A mechanism for **extending the format** by declaring new column or table types.

While the general rules apply to all kinds of data, the current version of SBtab is tailored for describing the structure of biochemical network models and the biochemical quantities therein. This is reflected by the table types defined in Table 1.

**Colour highlighting and predefined terms** In the examples shown in this text, predefined SBtab entries are highlighted in colours. This is just for convenience and is not a part of the SBtab format. `Table types` and `Column types` defined by the SBtab format are listed in Table 1. `Shortnames` can be chosen *ad hoc* by the user; each of them needs to be defined by a table row. Shortnames have to be unique and consistent within a document, but may differ between documents. `Reserved names` are predefined in SBtab for recurrent mathematical expressions like "mean value". `Official names`, like the names used for databases, are defined by some other authority. `Free text` and other text including database IDs, numerical values, mathematical brackets, and operators is written in black.

## 2.2 SBtab tables and SBtab documents

**General table structure** An SBtab document consists of one or several tables that refer to a common model or related data sets. All tables must use a common list of shortnames. For instance, a `Compound` table contains the column `!Compound`, and the elements from this column define compound shortnames to be used in the other tables. Several tables in a document may have the same type, but their table names (attribute `TableName`) must be unique.

**Declaration row containing the table attributes** The top left field contains the table header, starting with `!!SBtab` and followed by the table attributes in the syntax *attribute name=´attribute value´*, separated by whitespaces. Mandatory attributes are `TableType` and `TableName`.

**Column headers and definition table** The second row contains the column headers. Columns whose headers start with a `!` are treated as SBtab columns and must adhere to the SBtab rules. Other columns can contain arbitrary content. SBtab has a number of predefined table types that can hold different kinds of data. Each table type has a number of mandatory or optional columns with specific properties. An overview is given below and in the appendix. However, users can also define their own table types and corresponding columns. This definition must be provided by the user in the form of a special `Definition` table (as described below).

**Column with unique keys** By default, any SBtab table must start with a column matching its table type (e.g., a table of type `Quantity` must start with a column `Quantity`) and containing shortnames that serve as unique identifiers for the table elements. If a table does not have such a column with unique key, this should be marked by setting the attribute `UniqueKey='False'` in the declaration row of the table. The attribute is set to `True` by default.

**Completeness** To interpret the contents of a single table, other tables (e.g. describing shortnames) may be required. If a table does not require any other tables, we call it "complete". A document is complete if all names are defined, i.e. no unspecified information is required to interpret its contents. If a single table or a document are incomplete, the undefined names have to be known by the software, and an exchange with other software tools is likely to fail. If a table or document contains two elements, and there is no explicit information implying that they describe the same things, it is assumed that they describe different things.

| Name | Contents | Usage |
|---|---|---|
| `Compound` | Names, IDs, properties of compounds | model structure |
| `Enzyme` | Names, properties of enzymes | model structure |
| `Protein` | Names, properties of proteins | model structure |
| `Gene` | Names, properties of genes | model structure |
| `Regulator` | Names, properties of gene regulators | model structure |
| `Compartment` | Names and IDs of compartments | model structure |
| `Reaction` | Chemical reactions | model structure |
| `Quantity` | Individual data for model parameters | quantitative data |
| `QuantityMatrix` | Data matrices | quantitative data |
| `Relation` | Relations between different compounds | model structure |
| `Definition` | Define custom column types, etc. | customise SBtab |

Table 1: Overview of table types predefined in SBtab.

**Conventions for spreadsheet files** To ensure consistency between spreadsheet files, we propose a number of rules for good practice:

- **UTF8 encoding** If possible, the UTF8 encoding should be chosen.
- **Documents** In character-separated text files (.csv or .tsv), a document can either be stored in several files with the filenames *basename_tablename.extension*, or tables are concatenated vertically, each preceded by a declaration row (starting with `!!`), and stored in a single table file.
- **Delimiters in .csv or .tsv files** In character-separated files, irrespective of the extension (.csv or .tsv), it is assumed by default that the delimiters are tabulators. However, other delimiters (comma or semicolon) are accepted by the parser as well.
- **Special characters** If table cells contain special characters that are also used as cell delimiters (e.g. commas), the file must be provided in a form that excludes ambiguities (e.g. in the case of a comma-separated table containing commas with its fields, all cells must additionally be marked by quotation marks (`".."`).

**Filenames** The SBtab format as such does not impose any restrictions on filenames, nor does it require a specific filename extension. SBtab files stored as excel sheets, for instance, will have the extension `.xls`. However, the SBtab online tools (and the python programs behind it) have a certain convention for filenames and filename extensions. When an SBtab document is exported to several delimited text files, the filenames will be chosen according to the scheme `[SBTAB DOCUMENT NAME]_[TABLE TYPE].csv` or, in case of ambiguities `[SBTAB DOCUMENT NAME]_[TABLE TYPE]_[TABLE NAME].csv`.

**Filename extensions** Regarding filename extensions, the python implementation of SBtab supports comma-separated and tab-separated tables, as well as excel spreadsheet files (xls). By default, the python code exports tab-separated files and uses the filename extension `.csv`. This is a convention supported by LibreOffice, but may lead to conflicts in other cases. Some tools require extensions like `.tab` (excel) or `.tsv` (e.g., the formatting option in github). We do not use `.tsv` in this case, because this is supported neither by excel nor by LibreOffice. In case of conflicts, users may have to simply rename their files. When importing a table, the code tries to determine whether commas or tabs are used as delimiters. When using commas as delimiters, users have to make sure that no commas are used elsewhere in the table (or that all elements are given in double quotes).

## 2.3 Names of biochemical elements

**Names and identifiers of model elements** In the following, compounds, enzymes, genes, genetic regulators, and compartments will be called "biochemical entities". "Biochemical elements" comprises, in addition, reactions and biochemical quantities. Biochemical elements can be described by shortnames, official names, or database identifiers (IDs). The shortnames have to be declared within the SBtab document and have to satisfy syntactic rules. Each table starts with a column of the same name, containing the shortnames. Shortnames, the arbitrary element names used in a data set or model, must be unique, i.e. declared

only once in a document; they must start with a letter and may not contain spaces or the special characters ":", ".". In columns containing database IDs, the column name (`!Identifiers:`*Identifiers*) specifies the database by a name (to be used in column names, IDs etc.) and an URI. We suggest to use preferably the databases listed in the Miriam file (see Table 16). Sometimes, elements may be characterised redundantly: e.g. the reaction catalysed by an enzyme, given in an `Enzyme` table, can be given by both shortname and database ID. In case of conflict, the information derived from the shortname (i.e. the database ID listed in the `Reaction` table) has higher priority.

**Naming and specification of biological entities** Tables of the types `Compound`, `Enzyme`, `Gene`, `Regulator`, or `Compartment` are called "entity tables". The biochemical meaning of the entities can be declared by different columns:

- `!Name` contains official names (it is good practice to use names from the suggested databases). Several names can be listed in one field, separated by "|". To declare from which database a name has been taken, the name can also be written as *DB*:*name*.
- `!Identifiers:`*Identifiers* contains IDs from a specified database. Annotations with database IDs follow the scheme defined by Identifiers.org [5] (data collection and ID).

**Localised compounds** If a compound, enzyme, or genetic regulator is localised in a compartment, the corresponding localised entity can be denoted by *compound*[*compartment*] with square brackets, where *compound* and *compartment* are the shortnames or IDs of the compound and the compartment used in the model. If a model contains several compartments, tools should treat the first compartment in the `Compartment` table as the standard compartment. The standard compartment will be used by default for all compounds that are not explicitly assigned to compartments.

## 2.4 Annotating biochemical elements with database identifiers

Biochemical elements are annotated with database IDs listed in special identifier columns. An `Identifiers` column contains annotations from one web resource, at most one annotation per element, and without qualifiers. The column item and the referenced ID are assumed to be linked by an "is" relationship (and not, for instance, "version of", which can exist in SBML annotations). A table can contain several `Identifiers` columns, which must refer to different data resources.

| !!SBtab | TableName='Ex 7 - Compound' | TableType='Compound' | |
|---|---|---|---|
| !Compound | !Identifiers:obo.chebi | !Identifiers:kegg.compound | ... |
| water | CHEBI:15377 | C00001 | ... |
| ATP | CHEBI:15422 | C00002 | ... |
| phosphate | CHEBI:18367 | | ... |

To translate an element like `CHEBI:16865` into a valid Identifiers.org URI, `http://identifiers.org/` is concatenated with the data collection mentioned after `!Identifiers:` in the header (e.g. `obo.chebi`) and with the column item, separated by a slash[2]. For instance, the first annotation entry in the table above would be resolved to `http://identifiers.org/obo.chebi/CHEBI:15377`.

## 2.5 Syntax for reaction formulae

Chemical reactions can be described by reaction formulae (column `!SumFormula` in table `Reaction`; specifying the reactants, their stoichiometric coefficients, and possibly their localisation). The reaction arrow is denoted by <=>. Stoichiometric coefficients refer to substance amounts, not concentrations (this matters in the case of transport reactions). Stoichiometric coefficients of 1 are omitted; general stoichiometric coefficients, given by letters (e.g. `n`) are not allowed. If possible, the reaction formula should represent the actual stoichiometries experienced by the enzyme (i.e. `A <=> 2 B` rather than `0.5 A <=> B`). Substrates and products are given by shortnames, which must be defined in a `Compound` table. The

[2]The elements from the column have to be translated into a URN-encoded form (as described in the URN specification): for instance, the colon in the identifier `CHEBI:16865` has to be replaced by the string "`%3A`" to create the URN `obo.chebi:CHEBI%3A16865`.

order of substrates and the order of products are arbitrary; however, comparison of formulae is eased by using an alphabetical order. The localisation in compartments can be denoted as follows:

- Reaction in the default compartment: `A + 2 B <=> C + D`
- Transport reaction: `A[comp1] + 2 B[comp1] <=> C[comp2] + D[comp2]`

In the example, `A`, `B`, `C`, and `D` are compound shortnames, and `comp1` and `comp2` are compartment shortnames. The reversibility of reactions is not given by the sum formula, but by an extra column `!IsReversible` in the `Reaction` table.

# 3 Overview of predefined table types

SBtab predefines a number of table types with specific properties. An overview is given in Table 1. The table types `Compound`, `Enzyme`, `Gene`, `Regulator`, `Compartment`, and `Reaction` describe model structures, the table types `Quantity`, `QuantityMatrix`, and `Relation` are used for quantitative data.

## 3.1 Tables for biochemical network structures

As in example 1 (in the introduction section), biochemical networks consist of biochemical entities (e.g. metabolites or proteins) and reactions or interactions between them. The tables describing these entities (table types `Reaction`, `Compound`, `Compartment`, `Enzyme`, `Regulator`, and `Gene`) have to satisfy the following rules.

- **Entities** In tables describing biochemical entities (`Compound`, `Enzyme`, `Gene`, `Regulator`, `Compartment`), each row has to contain (i) a shortname as the primary key (in the column `!Compound`, `!Enzyme`, etc.) and (ii) at least one entry specifying the entity, like `!Name` or `!Identifiers:`*DB*. If a column shares the type of the table (e.g. a Compound column in a Compound table), it can be considered a primary key, that is, its elements should be unique and it should appear as the first column in the table. Optional columns - which may depend on the kinds of entities - are listed in Table B.2.
- **Reactions** A `Reaction` table lists chemical reactions, possibly with information about the corresponding enzymes, their kinetic laws, and their genetic regulation. It must contain at least one of the following columns: `!SumFormula`, `!Identifiers:`*DB*; optional columns are listed in Table 11. For an example, see example 1 in the introduction.
- **Enzymes, genes, and regulators** The connection between chemical reactions, the enzymes catalysing the reactions, and the genes coding for the enzymes can be complicated, but in many cases, there is a one-to-one relationship. In SBtab, there are different ways to express this relationship. Information about enzymes or genes and their regulation can be stored in a `Reaction` table if there is a one-to-one relationship between reactions, enzymes, and possibly genes. Otherwise, it is stored in separate tables `Enzyme` and `Gene` and the tables are interlinked *via* the columns `!Enzyme` (in table `Reaction`) and `!Gene` (in table `Enzyme`) or `!TargetReaction` (in an `Enzyme` table) and `!GeneProduct` (in a `Gene` table).

## 3.2 Table type `Quantity` for biochemical parameters

Numerical data (e.g. for time series or kinetic parameters) can be stored in tables and be linked to model elements via the latters' shortnames. There are two different table types for numerical data. Tables of type `Quantity` describe individual physical or biochemical quantities, for instance, kinetic parameters in a network model. These quantities can be linked to one entity, one reaction or enzyme, or both. If a quantity table contains several values for the same quantity, they appear in separate rows (for possible descriptions of provenance, see Table 10).

Tables of type `Quantity` describe single physical or biochemical quantities (e.g. individual kinetic constants). A quantity is defined by a type, a unit, possibly biochemical entities to which it refers, possibly a localisation, and possibly experimental or physical conditions. The columns contain the defining properties

(e.g. unit, conditions, etc.) and their values. Quantities can refer to a compound, an enzyme or reaction, or a combination of them. For instance, a concentration refers to a substance, while a $k^{\mathrm{M}}$ value refers to a metabolite and an enzyme. If there is a one-to-one relationship between reactions and enzymes, the $k^{\mathrm{M}}$ value can also be assigned to a compound/reaction pair or a compound/enzyme pair. Let us consider again example 2:

| !!SBtab | TableName='Ex 8 - Quantity' | TableType='Quantity' | | | |
|---|---|---|---|---|---|
| !Quantity | !QuantityType | !Reaction:Identifiers:kegg.reaction | !Compound:Identifiers:kegg.compound | !Value | !Unit |
| keq_R1 | equilibrium constant | R01061 | | 0.0984 | dimensionless |
| kmc_R1_C1 | Michaelis constant | R01061 | C00003 | 0.96 | mM |
| kic_R1_C1 | inhibition constant | R01070 | C00111 | 0.13 | mM |
| con_C1 | concentration | | C00118 | 0.203 | mM |

To specify the parameters of a model, we refer to `Reaction` and `Compound` elements by shortnames rather than by resource IDs. In this form, the above example becomes

| !!SBtab | TableName='Ex 9 - Quantity' | TableType='Quantity' | | | |
|---|---|---|---|---|---|
| !Quantity | !SBO:Identifiers:obo.sbo | !Reaction | !Compound | !Value | !Unit |
| kcrf_R1 | SBO:0000320 | R1 | | 200.0 | 1/s |
| keq_R1 | SBO:0000281 | R1 | | 0.0984 | dimensionless |
| kmc_R1_C1 | SBO:0000027 | R1 | C1 | 0.96 | mM |
| kic_R1_C2 | SBO:0000261 | R1 | C2 | 0.13 | mM |
| con_C3 | SBO:0000196 | | C3 | 0.203 | mM |

This example shows that quantity types can be specified by identifiers from the Systems Biology Ontology (SBO) in a column `!SBO:Identifiers:obo.sbo`.

A `Quantity` table can also store state-dependent quantities like concentrations, expression levels, or fluxes, like in the following example.

| !!SBtab | TableName='Ex 12 - Quantity' | TableType='Quantity' | | |
|---|---|---|---|---|
| !Quantity | !Compound | !Condition | !SBO:concentration | !Unit |
| con_C1_wt | C1 | wildtype | 0.2 | mM |
| con_C2_wt | C2 | wildtype | 1 | mM |
| con_C3_wt | C3 | wildtype | 0.1 | mM |
| con_C1_mu | C1 | mutant | 0.1 | mM |
| con_C2_mu | C2 | mutant | 0.5 | mM |
| con_C3_mu | C3 | mutant | 0.1 | mM |

## 3.3 Table type `QuantityMatrix` for data matrices

Biological data often have the form of matrices. As an example, consider a small 2×2 matrix containing metabolite concentrations for two time points and two metabolites. It can be expressed by the following SBtab table.

| !!SBtab | TableType='QuantityMatrix' | TableName='Ex 13 - Metabolomics data' UniqueKey='False' |
|---|---|---|
| !Time | Glucose | Fructose |
| 0.0 | 1.1 | 1.4 |
| 0.5 | 1.2 | 0.9 |

The headers of the data columns are not defined headers starting with "!", but simple strings. Therefore, they are not formally controlled by SBtab. Annotating these columns, e.g., by adding ChEBI Identifiers to specify the metabolites, is not directly possible. Moreover, the time points have no keys to which other tables could refer. An alternative solution looks as follows:

| !!SBtab | TableType='QuantityMatrix' | TableName='Ex 14 - Metabolomics data' | UniqueKey='False' |
|---|---|---|---|
| !TimePoint | !Time | >Measurement:Glucose | >Measurement:Fructose |
| T0 | 0.0 | 1.1 | 1.4 |
| T1 | 0.5 | 1.2 | 0.9 |

Here, the column headers are controlled and point to rows of another table with table name "Measurement", in which the ChEBI Identifers are given:

| !!SBtab | TableType='Quantity' | TableName='Ex 15 - Measurement' | UniqueKey='False' |
|---|---|---|---|
| !Compound | !Identifiers:obo.chebi | !QuantityType | !Unit |
| Glucose | CHEBI:17234 | concentration | mM |
| Fructose | CHEBI:15824 | concentration | mM |

Now let us consider data tables in which time points are represented by columns. A similar scheme can be used in this case. The first, simple version would read:

| !!SBtab | TableType='QuantityMatrix' | TableName='Ex 16 - Metabolomics data' | UniqueKey='False' |
|---|---|---|---|
| !Compound | !Identifiers:obo.chebi | t = 0 s | t = 0.5 s |
| Glucose | CHEBI:17234 | 1.1 | 1.2 |
| Fructose | CHEBI:15824 | 1.4 | 0.9 |

Here, it would obviously be good to store time point and time unit separately instead of merging them in the column header. This can be realised as follows:

| !!SBtab | TableType='QuantityMatrix' | TableName='Ex 17 - Metabolomics data' | UniqueKey='False' |
|---|---|---|---|
| !Compound | !Identifiers:obo.chebi | >TimePoint:t0 | >TimePoint:t1 |
| Glucose | CHEBI:17234 | 1.1 | 1.2 |
| Fructose | CHEBI:15824 | 1.4 | 0.9 |

with an extra table

| !!SBtab | TableType='Quantity' | TableName='Ex 18 - TimePoint' UniqueKey='False' |
|---|---|---|
| !TimePoint | !Time | !Unit |
| t0 | 0 | s |
| t1 | 0.5 | s |

## 3.4 The table type `Relation` for pairwise relations

The table type `Relation` is used to define pairwise links between objects. Each link ("relationship") can have a type and a numerical value. A `Relation` table can, for instance, be used to define a directed graph (by listing the edges between nodes of one type) or a gene regulatory network (by listing the actions of transcription factors on gene promoters). In particular, `Relation` tables can be used to link SBtab elements between tables and, thus, to create SBtab documents that have the form of a relational database.

| !!SBtab | TableName='Ex 19 - LittleGraph' | TableType='Relation' | UniqueKey='False' |
|---|---|---|---|
| !From | !To | !Relation | !Value |
| A | A | regulates | 1 |
| A | B | regulates | -1 |
| B | A | regulates | 1 |
| B | C | regulates | 2 |
| C | D | regulates | 1 |

## 3.5 Table type `Definition` for customising the SBtab format

Users can define their own table types and corresponding columns. For usage in the online tools or in the python code, this definition can be provided by the user in the form of a special `Definition` table. The default table (containing the predefined table and column types) is available on the SBtab website. Note that, when using a new Definition table, the predefined Definition table will be completely overridden, so any tables and columns to be used (also the predefined ones) must be listed in the new table. The typical format of a `Definition` table is shown below.

| !!SBtab | TableType='Definition' | TableName='Ex 20 - Def' | | |
|---|---|---|---|---|
| !Component | !ComponentType | !IsPartOf | !Format | !Description |
| SBML:reaction:id | Column | Reaction | String | SBML ID of reaction |
| SumFormula | Column | Reaction | String | Reaction sum formula |
| Enzyme | Column | Reaction | String | Enzyme catalysing the reaction |
| ... | ... | ... | ... | .. |

The `Format` column defines which type of entries a column can contain. Possibilities are `String`, `Shortname` (name of SBtab element, as defined in one of the SBtab tables), `Number` (integer or float

in usual formats, or complex numbers like `1 + i 3`), or `Boolean` (with possible values `True` and `False`, or alternatively 1 and 0). More specific string formats (e.g., for reaction sum formulae) are currently not formally defined, but can be mentioned in the `Description` column.

# 4 Conversion between SBtab and SBML

SBML (Systems Biology Markup Language) models can be converted into SBtab documents and vice versa. Depending on the content of the SBML model, the SBtab files can comprise table types `Reaction`, `Compound`, `Compartment`, `Quantity`, `Events`, and `Rules`. Likewise, these SBtab table types can be converted into an SBML (Level 2, Version 4) model. The conversion to SBML, however, requires at least either a `Reaction` or `Compound` SBtab.

The conversion from an SBML model file to SBtab translates the structural and temporal information of the model into corresponding SBtab table files. The (i) `Reaction` SBtab contains a list of the reactions of the SBML file, including their sum formula, kinetic laws, irreversibility, annotations, and more. Note that the SBML modifiers of a reaction (e.g. enzymes) cannot be identified as inhibitor or stimulator if they are not assigned an SBO Term within the SBML code. If this is not the case, they will only be exported to SBtab as modifiers without regulatory information. All species from the model can be found in the (ii) `Compound` SBtab. Their location, charge, annotations, and more are provided in the SBtab. Analogously, a (iii) `Compartment` holds all structural information of the cellular compartments. The (iv) `Quantity` SBtab file lists all parameters that are part of the model. Also their numerical values and units will be provided. The parameters can appear as either local or global variables in the SBML code; this information will be transferred to SBtab as well. (v) `Events` can be an important part of SBML models; they indicate e.g. concentration changes or stress applications at certain time points. They too are translated into an SBtab file. Finally, (vi) `rules` are exported from SBML to an SBtab Rule table. Rules can comprise assignment rules, algebraic rules, and rate rules. Rule formulas and units are part of the conversion as well.

In the conversion from SBtab to SBML, `Compound` entries in SBtab correspond to `species` elements in SBML. By default, the unique keys in the `Compound` and `Reaction` SBtab are used as `id` attributes of the SBML elements. If SBML IDs are directly specified within SBtab (in the columns `SBML:reaction:id`, `SBML:species:id`, `SBML:parameter:id`, `SBML:reaction:parameter:id`, etc), these will be used instead. Rate laws from the SBML code are stored in SBtab as strings within a `KineticLaw` column. Note that the rate laws are not checked for their validity. It is up to the user to assure the correctness of the rate laws. If they are erroneous, this leads to invalid SBML output. An automatic parser of rate laws including checks of validity is planned for future versions of SBtab. `Regulator` entries in SBtab correspond to `modifier` elements in SBML; multiple regulators can be described by a regulation formula (in the `Regulator` column): regulators are separated by a "|" symbol, while the sign of regulation can be denoted by + or -. For an enzyme allosterically inhibited by ATP and activated by ADP, the formula reads `-ATP|+Pyruvate` or `ATP|ADP` where inhibition and activition remain unspecified. Also rate rules and assignment rules can be converted from SBtab to SBML. Note that, just like for kinetic rate laws, these rules do not underlie constraints of validity. It is up to the user to ensure their correctness before conversion to SBML. Finally, SBtab is able to provide lists of events for the SBML file. This includes the event assignments, triggers, delays, and more. For all aforementioned SBML elements, annotations are automatically translated from the SBtab to the SBML file, if they adhere to the correct syntax.

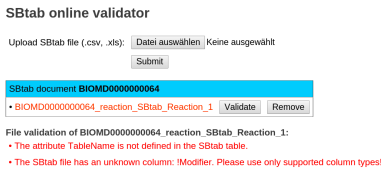
The entries of `Quantity` tables can be inserted into SBML models or be extracted from them. By default, SBtab quantities referring to a reaction will become local reaction parameters in SBML, while other quantities become global parameters. The element of the `!Quantity` column will be used as SBML element ID unless it is overridden by the (optional) column `!SBML:parameter:id` (for global parameters) or `!SBML:reaction:parameter:id` (for local reaction parameters). Naming conventions for kinetic constants are given in [6], supplementary material Table A.5. Quantities that describe initial species amounts, initial species concentrations, or compartment sizes will be translated into the corresponding SBML element attributes.

There are still limitations to the conversion of SBtab and SBML. So far, the conversion does not include element notes and function definitions. These issues are planned to be solved in future versions of SBtab.

# 5 SBtab tools

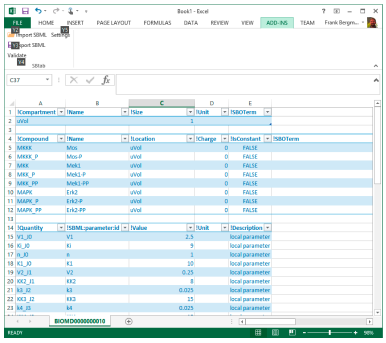
To simplify the usage of SBtab, we provide several online tools at `www.sbtab.net`.

**1. Online validator for SBtab files.** The online validator tool checks whether SBtab files (in .csv or .xls format) adhere to the SBtab conventions introduced in this manuscript. If a problem is identified by the validator, an instruction on how to fix the problem is provided. The validation is based on the SBtab table definitions found in the `Definition` table.




**2. Online SBtab ↔ SBML converter** The online conversion tool can create SBtab files from SBML models and vice versa. For the conversion from SBtab to SBML, it has to be assured that at least an SBtab table of type `Reaction` or `Compound` is provided. As additional information, the following SBtab table types can be used for the conversion to SBML:`Compartment`, `Quantity`, `Events`, and `Rules`. All information comprised in these SBtab tables can be converted to the SBML structure, as long as they are adhering to the correct syntax. Therefore, it is recommended to validate the SBtab files with the online validator before recruiting them for a conversion to SBML. The generated SBtab files can be displayed online as HTML tables. If annotations are correctly provided, they will link to the web resource. For the conversion it is recommended to use SBML Level 2, Version 4, or higher. The details on the conversions can be read in Chapter 4.


**3. MS Excel Add-in** The described validator and converter functions can also be attained with an add-in for Microsoft Excel. It can be retrieved from the SBtab Github Repository and installed with a Windows Installer Package. The prerequisites for the installation of the add-in are (i) Windows Vista or higher, (ii) Microsoft Office 2010 or higher, (iii) Microsoft .NET Framework 4.5 (full) or higher, and Microsoft Visual Studio 2010 Tools for Office Runtime (VSTO). The latter two can be downloaded directly from Microsoft.



**Python parser for SBtab files.** In addition, we provide a SBtab parser written in Python. It uses the Python package `tablib` to import SBtab files and provides different functions for editing the data and for directly accessing them. These features are important for the embedding of the SBtab file parser into software projects. The common operations for manipulating SBtab files contain:

1. Extracting characteristic table information (type, name, etc.).
2. Addition of rows and columns to the SBtab table.
3. Editing and export of the table content in rows, columns, and single entries. An export as a Python dictionary is also possible, to ensure easy access to the data for python programmers.
4. Switching of columns and rows in the table (matrix transposition). As some data are stored conveniently in transposed spreadsheets, some tables need to be transposed to have better access to its content.
5. Duplicate SBtab objects.
6. Writing SBtab files to the hard disk.

## Acknowledgements

The authors thank Dagmar Waltemath, Hans-Michael Kaltenbach, Dirk Wiesenthal, Jannis Uhlendorf, Anne Goelzer, Jörg Büscher, Avi Flamholz, Elad Noor, Edda Klipp, Frank Bergmann, Phillipp Schmidt, and Matthias König for contributing to this proposal. This work was funded by the German Research Foundation (LI 1676/2-1), the European Commission (projects BaSysBio and UNICELLSYS), and the German Federal Ministry of Education and Research (project OncoPath).

## References

[1] M. Kanehisa, S. Goto, S. Kawashima S, and A. Nakaya. The KEGG databases at genomenet. *Nucleic Acids Research*, 30:42–46, 2002.

[2] C. Laibe and N. Le Novère. MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Systems Biology*, 1(1):58, 2007.

[3] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I.I. Goryanin, W.J. Hedley, T.J. Hodgman, J.H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada T J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and the SBML Forum. The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[4] M. Courtot et al. Controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, 7(543), 2011.

[5] N. Juty, N. Le Novère, and C. Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research*, 40:D580–D586, 2012.

[6] W. Liebermeister, J. Uhlendorf, and E. Klipp. Modular rate laws for enzymatic reactions: thermodynamics, elasticities, and implementation. *Bioinformatics*, 26(12):1528–1534, 2010.

[7] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[8] N. Le Novère, A. Finney, M. Hucka, U.S. Bhalla, F. Campagne, J. Collado-Vides, E.J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J.L. Snoep, H.D. Spence, and B.L. Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotech.*, 23(12):1509–1515, Dec 2005.

# A Summary of SBtab rules

We summarise the most important conventions implemented by SBtab:

- **Shortnames** Model elements (e.g. compounds) are referred to by shortnames, which are defined in the corresponding table (e.g. `Compound` for compounds) . Shortnames must be unique within an SBtab document. The first column of each table shares the name of the table type (e.g. column `!Compound` in table type `Compound`) and contains the shortnames, which serve as primary keys for this table and must therefore be unique. If a table does not contain such a unique key column, this must be declared by setting the table attribute `UniqueKey='False'` (this can be the case for tables of type `QuantityMatrix`, for instance).
- **Order of columns** The allowed column types depend on the table type, but their order is arbitrary. The only exception is the first column, which contains the shortnames (acting as keys for this table) and whose name corresponds to the table type. However, it is good practice to sort the columns by importance and to arrange related columns next to each other (e.g. placing a column `Value` next to a column `Unit`).
- **ASCII Characters** The table fields contain only plain text. The format is case-sensitive, but the choice of fonts (bold, italic) does not play a role. Double quotes should not be used.
- **Decimal points** To simplify parsing, we recommend to use decimal points (instead of decimal commas).
- **Table types and column names** Table types and their possible columns are defined in appendix B. Column names may not contain any special characters or white spaces (parsers should ignore these characters).
- **Comment lines** Table lines starting with a "%" character contain comments and are ignored during parsing.
- **Comments and references** Additional information about table elements can be stored in the optional columns `!Comment`, `!Reference`, `!Reference:Identifiers:pubmed`, and `!ReferenceDOI`, which can appear in all tables.
- **Unrecognised table or columns** Columns with unknown headers (not starting with `!`), or unrecognised header starting with `!` may appear in SBtab tables. They can be used, but are not supported by the parser. The use of undefined columns is inadviseable.
- **Declaration line** The first line, starting with `!!SBtab` must declare at least the attributes: `TableType`, `TableName`, and possibly the properties `SBtabVersion` (for SBtab version used) and `Document`. The entries can be separated by whitespaces or be given in separate fields of the declaration line.
- **Identifiers** Identifiers for compounds, compartments etc. can be specified in columns with a header "*ElementType*:`Identifiers:`*DB*").
- **Missing elements** If an element is missing, the table field is left empty. Missing numerical values can also be indicated by non-numerical elements like `?` or `na` (for "not available"). Mandatory fields must not be empty.
- **Formulae** Reaction sum formulae must be written in a special format explained below.
- **Reserved names** In the SBtab format, there are reserved names for (i) table types (marked by colours in this text); (ii) column names; (iii) types of biological elements (see Table 17); and (iv) types of biochemical quantities or mathematical terms (e.g. `Mean`) for them (see Table 18), and physical units.
- **Physical units** In SBtab, it is recommended to use the units listed in the SBML specification (see sbml.org/Documents/Specifications)[3]. As good practice, derived units (e.g. `kJ/mol`) and reciprocal units (e.g. `1/s`) should be given in the simplest possible form, in necessary using multiplication, division, exponentials, and round brackets (e.g. `gram/m^3`).

[3]The following units are supported by SBML: `ampere`, `gram`, `katal`, `metre`, `second`, `watt`, `becquerel`, `gray`, `kelvin`, `mole`, `siemens`, `weber`, `candela`, `henry`, `kilogram`, `newton`, `sievert`, `coulomb`, `hertz`, `Litre`, `ohm`, `steradian`, `dimensionless`, `item`, `lumen`, `pascal`, `tesla`, `farad`, `joule`, `Lux`, `radian`. Orders of magnitude can be denoted by `k`, `M`, `c`, `m`, `mu`, `n`, `p`, `f` for Kilo, Mega, Centi, Milli, Micro, Nano, Pico, Femto. If a parameter is dimensionless, it has to be annotated as `dimensionless`.

# B Overview of table types

## B.1 Document and table attributes and general column types

### Table attributes

| Name | Type | Format | Mandatory | Content |
|---|---|---|---|---|
| TableType | text | string | ✓ | Table type (as defined in definition table) |
| TableName | text | string | ✓ | Table name |
| SBtabVersion | text | string | | SBtab version number |
| Document | text | string | | SBtab document name |
| UniqueKey | text | Boolean | | Requirement of a unique key identifier |
| ReferenceDescription | text | string | | Name of reference description |
| Document | text | string | | Document name |
| ReferenceCitation | text | string | | Citation, unique identifier, unambiguous URL |
| ModelCreators | text | string | | Name and contact information for model creators |
| ModelCreationTime | text | string | | Date and time of model creation and last modification |
| TermsOfDistribution | text | string | | Terms of distribution |

Table 2: Possible table attributes (to appear in declaration row). The attributes in the lower part would be necessary for MIRIAM compliance. If ReferenceCitation contains a pubmed Id, the attribute ReferenceCitation:Identifiers:pubmed should be used instead. ReferenceCitation should also identify the authors of the model.

### All table types

| Name | Type | Format | Content |
|---|---|---|---|
| !Description | text | string | Description of the row element |
| !Comment | text | string | Comment |
| !ReferenceName | text | string | Reference title, authors, etc. (as free text) |
| !Reference:Identifiers:pubmed | text | string | Reference PubMed ID |
| !ReferenceDOI | text | string | Reference DOI |

Table 3: Columns that can appear in all tables

### All entity and reaction tables

| Name | Type | Format | Content |
|---|---|---|---|
| !Name | text | string | Entity name |
| !Identifiers:*DataCollection* | resource ID | string | Entity ID |
| !MiriamAnnotations | annotation | string | Entity ID (JSON string) |
| !Type | text | string | Biochemical type of entity (examples see Table 17) |
| !Symbol | text | string | Short symbol (e.g., gene symbol) |
| !PositionX | number | float | x coordinate for graphical display |
| !PositionY | number | float | y coordinate for graphical display |

Table 4: Columns that can appear in all entity (i.e. Compound, Enzyme, Gene, Regulator, and Compartment) and Reaction tables.

## B.2 Predefined table types

### Compound

| Name | Type | Format | Content |
|---|---|---|---|
| !Compound | shortname | string | Compound shortname |
| !SBML:species:id | SBML element ID | string | SBML Species ID of the entity |
| !SBML:speciestype:id | SBML element ID | string | SBML SpeciesType ID of the entity |
| !InitialValue | number | float | Initial amount or concentration |
| !Unit | string | string | Unit for initial value |
| !Location | shortname | string | Compartment for localised entities |
| !State | shortname | string | State of the entity |
| !CompoundSumFormula | text | string | Chemical sum formula |
| !StructureFormula | text | string | Chemical structure formula |
| !Charge | number | integer | Electrical charge number |
| !Mass | number | float | Molecular mass |
| !Unit | text | string | Physical unit |
| !IsConstant | Boolean | Boolean | Substance with fixed concentrations |
| !EnzymeRole | shortname | string | Enzymatic activity |
| !RegulatorRole | shortname | string | Regulatory activity |

Table 5: Columns that can appear in Compound tables

### Enzyme

| Name | Type | Format | Content |
|---|---|---|---|
| !Enzyme | shortname | string | Enzyme shortname |
| !CatalysedReaction | shortname | string | Catalysed reaction |
| !KineticLaw:Name | name | string | Rate law (name as in SBO) |
| !KineticLaw:Identifiers.obo.sbo | shortname | string | Rate law SBO identifier |
| !Pathway | text | string | Pathway name (free text) |
| !Gene | shortname | string | Gene coding for enzyme (shortname) |
| !Gene:Name | string | string | Gene coding for enzyme (name) |
| !Gene:Symbol | string | string | Gene coding for enzyme (short symbol) |

Table 6: Columns that can appear in Enzyme tables

### Protein

| Name | Type | Format | Content |
|---|---|---|---|
| !Protein | shortname | string | Protein shortname |
| !Name | text | string | Protein name |
| !Symbol | string | string | Protein symbol |
| !Gene | shortname | string | Gene shortname |
| !Gene:Name | text | string | Gene name |
| !Gene:Symbol | string | string | Gene symbol |
| !Gene:LocusName | string | string | Gene locus name |
| !Mass | number | number | Protein mass |
| !Size | number | number | Protein size |

Table 7: Columns that can appear in Protein tables

### Gene

| Name | Type | Format | Content |
|---|---|---|---|
| !Gene | shortname | string | Gene shortname |
| !Name | text | string | Gene name |
| !Symbol | string | string | Gene symbol |
| !LocusName | string | string | Gene locus name |
| !GeneProduct | shortname | string | Gene product shortname |
| !GeneProduct:Name | string | string | Gene product name |
| !GeneProduct:Symbol | string | string | Gene product symbol |
| !GeneProduct:SBML:species:id | SBML element ID | string | SBML ID of protein |
| !Operon | shortname | string | Operon in which gene is located |

Table 8: Columns that can appear in Gene tables

### Regulator

| Name | Type | Format | Content |
|---|---|---|---|
| !Regulator | shortname | string | Regulator shortname |
| !State | shortname | string | State of the regulator |
| !TargetGene | shortname | string | Target gene |
| !TargetOperon | shortname | string | Target operon |
| !TargetPromoter | shortname | string | Target promoter |

Table 9: Columns that can appear in Regulator tables

### Compartment

| Name | Type | Format | Content |
|---|---|---|---|
| !Compartment | shortname | string | Compartment shortname |
| !Identifiers:obo.sbo | shortname | string | Compartment SBO term |
| !SBML:compartment:id | SBML element ID | string | SBML Compartment ID |
| !OuterCompartment | shortname | string | Surrounding compartment (short) |
| !OuterCompartment:Name | string | string | Surrounding compartment (name) |
| !OuterCompartment:SBML:compartment:id | SBML element ID | string | Surrounding compartment |
| !Size | number | float | Compartment size |
| !Unit | text | string | Physical unit |

Table 10: Columns that can appear in Compartment tables

### Reaction

| Name | Type | Format | Content |
|---|---|---|---|
| !Reaction | shortname | string | Reaction shortname |
| !SBML:reaction:id | SBML element ID | string | SBML Reaction ID |
| !SumFormula | SumFormula formula | string | Reaction sum formula |
| !Location | shortname | string | Compartment for localised reaction |
| !Enzyme | shortname | string | Enzyme catalysing the reaction |
| !Model | text | string | Model(s) in which reaction is involved |
| !Pathway | text | string | Pathway(s) in which reaction is involved |
| !SubreactionOf | shortname | string | Mark as subreaction of a (lumped) reaction |
| !IsComplete | Boolean | Boolean | Reaction formula includes all cofactors etc. |
| !IsReversible | Boolean | Boolean | Reaction should be treated as irreversible |
| !IsInEquilibrium | Boolean | Boolean | Reaction approximately in equilibrium |
| !IsExchangeReaction | Boolean | Boolean | Some reactants are left out |
| !Flux | number | float | Metabolic flux through the reaction |
| !IsNonEnzymatic | Boolean | Boolean | Non-catalysed reaction |
| !KineticLaw:Name | name | string | Rate law (name as in SBO) |
| !KineticLaw:Identifiers.obo.sbo | shortname | string | Rate law SBO identifier |
| !Gene | shortname | string | see table type Enzyme |
| !Gene:Symbol | string | string | see table type Enzyme |
| !Operon | shortname | string | see table type Gene |
| !Enzyme:Name | string | string | Name of enzyme |
| !Enzyme:Identifiers:ec-code | string | string | EC number of enzyme |
| !Enzyme:SBML:species:id | SBML element ID | string | SBML ID of enzyme |
| !Enzyme:SBML:parameter:id | SBML element ID | string | SBML ID of enzyme |
| !Enzyme:SBML:reaction:parameter:id | SBML element ID | string | SBML ID of enzyme |
| !BuildReaction | Boolean | Boolean | Includereaction in SBML model |
| !BuildEnzyme | Boolean | Boolean | Include enzyme in SBML model |
| !BuildEnzymeProduction | Boolean | Boolean | Describe enzyme production in SBML model |

Table 11: Columns that can appear in Reaction tables. The lower section lists, again, column types from Table B.2.

### Relation

| Name | Type | Format | Content |
|---|---|---|---|
| !Relation | shortname | string | Type of quantitative relationship |
| !From | shortname | string | Element at beginning of arrow |
| !To | shortname | string | Element at arrowhead |
| !IsSymmetric | Boolean | Boolean | Flag indicating non-symmetric relationships |
| !Value:*QuantityType* | number | float | Numerical value assigned to the relationship |

Table 12: Columns that can appear in Relation tables.

`Quantity`

| Name | Type | Format | Content |
|---|---|---|---|
| `!Quantity` | shortname | string | Quantity / SBML parameter ID |
| `!QuantityType` | shortname | string | Quantity type (e.g. from SBO) |
| *ValueType* | ValueType | string | Mathematical Term from table 15 |
| `!SBML:parameter:id` | SBML element ID | string | Parameter ID in SBML file |
| `!SBML:reaction:parameter:id` | SBML element ID | string | Parameter ID in SBML file |
| `!Unit` | text | string | Physical unit |
| `!Scale` | text | string | Scale (e.g. logarithm, see Table 15) |
| `!Provenance` | text | string | Name of data source (free text) |
| `!Condition` | text | string | experimental condition name (free text) |
| `!pH` | number | float | pH value in measurement |
| `!Temperature` | number | float | Temperature in measurement |
| `!Location` | shortname | string | Compartment (shortname) |
| `!Location:Name` | string | string | Compartment (name) |
| `!Location:SBML:compartment:id` | SBML element ID | string | SBML ID of compartment‘ |
| `!Compound` | shortname | string | Related compound (shortname) |
| `!Compound:Name` | string | string | Related compound (name) |
| `!Compound:Identifiers:`*DataCollection* | resource ID | string | Compound ID |
| `!Compound:SBML:species:id` | SBML element ID | string | SBML ID of compound |
| `!Reaction` | shortname | string | Related reaction (shortname) |
| `!Reaction:Name` | string | string | Related reaction (name) |
| `!Reaction:Identifiers:`*DataCollection* | resource ID | string | Reaction ID |
| `!Reaction:SBML:reaction:id` | SBML element ID | string | SBML ID of reaction |
| `!Enzyme` | shortname | string | Related enzyme (shortname) |
| `!Enzyme:Name` | string | string | Related enzyme (name) |
| `!Enzyme:Identifiers:`*DataCollection* | resource ID | string | Enzyme ID |
| `!Enzyme:SBML:species:id` | SBML element ID | string | SBML ID of enzyme |
| `!Enzyme:SBML:parameter:id` | SBML element ID | string | SBML ID of enzyme |
| `!Enzyme:SBML:reaction:parameter:id` | SBML element ID | string | SBML ID of enzyme |
| `!Protein` | shortname | string | Related enzyme (shortname) |
| `!Protein:Name` | string | string | Related enzyme (name) |
| `!Protein:Identifiers:`*DataCollection* | resource ID | string | Protein ID |
| `!Protein:SBML:species:id` | SBML element ID | string | SBML ID of enzyme |
| `!Protein:SBML:parameter:id` | SBML element ID | string | SBML ID of enzyme |
| `!Protein:SBML:reaction:parameter:id` | SBML element ID | string | SBML ID of enzyme |
| `!Gene` | shortname | string | Related gene |
| `!Organism` | shortname | string | Related organism |

Table 13: Columns for numerical values and experimental conditions in tables of type `Quantity`.

`Definition`

| Name | Type | Content |
|---|---|---|
| `!Component` | component name | Name of component (table, column, attribute to be defined) |
| `!ComponentType` | `Table, Column, Attribute` | Type of component |
| `!IsPartOf` | component name | name of parent component |
| `!Format` | String | Format |
| `!Description` | Text | Free text description of component |

Table 14: Columns that can appear in `Definition` tables.

# C Predefined terms and recommended controlled vocabularies

| ValueType | Type | Format | Meaning |
|---|---|---|---|
| Value | number | float | Simple value |
| Mean | number | float | Algebraic mean |
| Std | number | float (positive) | Standard deviation |
| Min | number | float | Lower bound |
| Max | number | float | Upper bound |
| Median | number | float | Median |
| GeometricMean | number | float | Geometric mean |
| Sign | sign | {+,-,0} | Sign |
| ProbDist | Free text | string | Prob. distribution |

| Scale | Meaning |
|---|---|
| Lin | Linear scale (no transformation) |
| Ln | Natural logarithm |
| Log2 | Dual logarithm |
| Log10 | Decadic logarithm |

Table 15: Terms for mathematical quantities and mathematical scales recommended for use in SBtab. Names of probability distributions can be, for instance, `Normal`, `Uniform`, `LogNormal`.

| Database | MIRIAM URN | Contents | URI |
|---|---|---|---|
| SBO | obo.sbo | Quantities, rate laws | www.ebi.ac.uk/sbo/ |
| CheBI | obo.chebi | Metabolites | www.ebi.ac.uk/chebi/ |
| Enzyme nomenclature | ec-code | Enzymes | www.ebi.ac.uk/IntEnz/ |
| KEGG Compound | kegg.compound | Compounds | www.genome.jp/KEGG/ |
| KEGG Reaction | kegg.reaction | Reactions | www.genome.jp/KEGG/ |
| KEGG Orthology | kegg.orthology | Genes | www.genome.jp/KEGG/ |
| UniProt | uniprot | Proteins | www.uniprot.org/ |
| SGD | sgd | Yeast gene loci | www.yeastgenome.org/ |
| Gene Ontology | obo.go | Compartments | www.geneontology.org/ |
| Taxonomy | taxonomy | Organisms | www.ncbi.nlm.nih.gov/Taxonomy/ |
| SGD | sgd | Yeast proteins | www.yeastgenome.org/ |

Table 16: A selection of databases to be used in SBtab. For the complete list, see the MIRIAM resources [2].

| **Physical entity types** | |
|---|---|
| protein complex | SBO:0000297 |
| messenger RNA | SBO:0000278 |
| ribonucleic acid | SBO:0000250 |
| deoxyribonucleic acid | SBO:0000251 |
| polypeptide chain | SBO:0000252 |
| polysaccharide | SBO:0000249 |
| metabolite | SBO:0000299 |
| macromolecular complex | SBO:0000296 |

| **Compartments** | |
|---|---|
| cell | GO:0005623 |
| extracellular space | GO:0005615 |
| membrane | GO:0001602 |
| cytosol | GO:0005829 |
| nucleus | GO:0005634 |
| mitochondrion | GO:0005739 |

Table 17: Examples of biochemical entity types (with Systems Biology Ontology identifiers [4]) and cell compartments (with Gene Ontology identifiers [7]).

# D A note on MIRIAM-compliant models

The MIRIAM rules for computational models [8] have been established to guarantee that published models contain complete and unambiguous information, and that results from the models can be verified. Note that MIRIAM-compliance also involves criteria that cannot be ensured by the file structure alone, but are related to how the model was made, and to the existence of a reference publication (which may or may not exist for a given SBtab file). (i) The encoded model structure must reflect the biological processes described by the reference description. (ii) The model must be instantiable in a simulation: all quantitative attributes must be defined, including initial conditions. (iii) When instantiated, the model must be able to reproduce all results given in the reference description within an epsilon (algorithms, round-up errors).

However, to allow users to satisfy some of the MIRIAM requirements, SBtab contains document attributes for information that is mandatory for MIRIAM-compliance. These must be given in the declaration line of the SBtab document in question, or in the declaration lines of at least one tables belonging to the document (i) ReferenceDescription (ii) DocumentName (iii) ReferenceCitation (complete citation, unique

| Name | SBO term | Default unit | Entities |
|---|---|---|---|
| standard Gibbs energy of formation | SBO:0000582 | kJ/mol | Compound |
| standard Gibbs energy of reaction | SBO:0000583 | kJ/mol | Compound |
| equilibrium constant | SBO:0000281 | variable | Reaction |
| forward maximal velocity | SBO:0000324 | mMol/s | Enzymatic reaction |
| reverse maximal velocity | SBO:0000325 | mMol/s | Enzymatic reaction |
| substrate catalytic rate constant | SBO:0000321 | 1/s | Enzymatic reaction |
| product catalytic rate constant | SBO:0000320 | 1/s | Enzymatic reaction |
| Michaelis constant | SBO:0000027 | mM | Enzyme, Compound |
| inhibitory constant | SBO:0000261 | mM | Enzyme, Compound |
| activition constant | SBO:0000363 | mM | Enzyme, Compound |
| Hill constant | SBO:0000190 | dimensionless | Compound, Reaction |
| concentration | SBO:0000196 | mM | Compound |
| biochemical potential | SBO:0000303 | kJ/mol | Compound |
| standard biochemical potential | SBO:0000463 | kJ/mol | Compound |
| rate of reaction (amount) | SBO:0000615 | M/s | Reaction |
| rate of reaction (concentration) | SBO:0000614 | mM/s | Reaction |
| Gibbs free energy of reaction | SBO:0000617 | kJ/mol | Reaction |
| standard Gibbs free energy of formation | SBO:0000582 | kJ/mol | Compound |
| standard Gibbs free energy of reaction | SBO:0000583 | kJ/mol | Compound |
| transformed standard Gibbs free energy of reaction | SBO:0000620 | kJ/mol | Reaction |
| transformed standard Gibbs free energy of formation | SBO:0000621 | kJ/mol | Compound |
| transformed Gibbs free energy of reaction | SBO:0000622 | kJ/mol | Reaction |
| thermodynamic temperature | SBO:0000147 | K | Location (optional) |
| ionic strength | SBO:0000623 | mM | Location (optional) |
| pH | SBO:0000304 | dimensionless | Location (optional) |

Table 18: A selection of quantity types to be used in SBtab in table types `Quantity`. The unit of equilibrium constants depends on the reaction stoichiometry. More quantities can be found in the Systems Biology Ontology [4].

identifier, unambiguous URL). The citation should identify the authors of the model. (iv) ModelCreators (name and contact information for model creators) (v) ModelCreationTime (The date and time of model creation and last modification) (vi) TermsOfDistribution (link to a precise statement about the terms of it's distribution).

**Selbständigkeitserklärung**

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel in Anspruch genommen habe. Ich versichere, dass diese Arbeit in dieser oder anderer Form keiner anderen Prüfungsbehörde vorgelegt wurde.

Berlin, den 5.1.2016

Timo Lubitz

## Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive mit (x11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.